VideoLT: Large-scale Long-tailed Video Recognition (Supplementary Material)

Xing Zhang[1*] Zuxuan Wu[2,3*] Zejia Weng[2] Huazhu Fu[4]
Jingjing Chen[2,3] Yu-Gang Jiang[2,3†] Larry Davis[5]
[1]Academy for Engineering and Technology, Fudan University,
[2]Shanghai Key Lab of Intel. Info. Processing, School of Computer Science, Fudan University,
[3] Shanghai Collaborative Innovation Center on Intelligent Visual Computing,
[4]Inception Institute of Artificial Intelligence, [5]University of Maryland

1. Overview

We provide more details about the paper in this supplementary material:

- Comparisons with existing video recognition datasets in Sec. 2.
- More details about annotations in Sec. 3.
- Extensive experiments on more long-tailed methods in Sec. 4.
- Details of state-of-the-art long-tailed methods in Sec. 5.
- Extensions of ablation studies in Sec. 6
- Definition of head, medium and tail classes in Sec. 7.
- Full taxonomy of VideoLT dataset in Sec. 8.

2. Comparisons with Existing Video Recognition Datasets

*To the best of our knowledge, **VideoLT** is the first "untrimmed" video recognition dataset which contains more than 1,000 manually defined classes.* We compare VideoLT with existing video datasets that developed for video recognition. See table 1 for details.

3. Annotation

To obtain high quality annotations, 35 annotators were hired, and each video was annotated by 3 annotators. In particular, after the collection of all videos using coarse labels defined by our taxonomy, each video was assigned to 3 annotators to check if the video content is relevant to its tags. A major voting is conducted—if the video is positive then it is added to the video pool, otherwise it is discarded. After filtering, we collected 256,218 videos with coarse labels.

4. Extensions with More Long-tailed Methods

To better illustrate the effectiveness of FrameStack, we conduct extensive experiments on square-root sampling [10] and two-stage methods [10], including τ-normalization, cRT and LWS. From table 2, experiments under Nonlinear model show that square-root sampling (SRS) leads to performance drop and two-stage methods are not effective under long-tailed video recognition scenario.

Dataset	Year	# Classes	# Videos	Untrimmed?	Manually Labeled?
Hollywood2 [15]	2009	12	3669	Yes	No
HMDB51 [12]	2011	51	6,766	No	Yes
UCF-101 [19]	2012	101	13,320	No	Yes
Sports1M [11]	2014	487	1,133,158	Yes	No
Charades [18]	2016	157	9,848	Yes	Yes
ActivityNet v1.3 [2]	2016	200	19,994	Yes	Yes
Youtube-8M [1]	2016	4,800	8,264,650	Yes	No
HACS [23]	2017	200	520,000	No	Yes
Moments in Time [16]	2017	339	1,000,000	No	Yes
Something-Something v1 [9]	2017	174	108,499	No	No
Something-Something v2 [14]	2018	174	220,847	No	Yes
Kinetics400 [6]	2017	400	306,245	No	Yes
Kinetics600 [4]	2018	600	495,547	No	Yes
Kinetics700 [5]	2019	700	650,317	No	Yes
VideoLT	-	1,004	256,218	Yes	Yes

Table 1. Comparison of video recognition datasets.

	ResNet-50						ResNet-101					
LT-Methods	Overall	[500,+∞) Head	[100,500) Medium	[0,100) Tail	Acc@1	Acc@5	Overall	[500,+∞) Head	[100,500) Medium	[0,100) Tail	Acc@1	Acc@5
baseline	0.499	0.675	0.553	0.376	0.650	0.828	0.516	0.687	0.568	0.396	0.663	0.837
SRS	0.486	0.655	0.539	0.365	0.638	0.822	0.516	0.680	0.568	0.399	0.662	0.836
cRT	0.498	0.674	0.553	0.375	0.651	0.825	0.515	0.687	0.568	0.395	0.662	0.836
t-normalized	0.499	0.675	0.553	0.376	0.649	0.825	0.515	0.687	0.568	0.396	0.661	0.835
LWS	0.499	0.675	0.553	0.376	0.650	0.828	0.515	0.687	0.568	0.396	0.663	0.837
FrameStack	**0.516**	**0.683**	**0.569**	**0.397**	**0.658**	**0.834**	**0.532**	**0.695**	**0.584**	**0.417**	**0.667**	**0.843**

Table 2. Comparing FrameStack with two-stage methods and square-root sampling.

5. Details of State-of-the-art Long-tailed Methods

Re-sampling methods:

- Class-balanced Sampling [10, 17]

  In a mini-batch, it takes a random class then it randomly samples a video, and thus videos from head and tail classes share the same probabilities to be chosen.

- Square-root Sampling [10]

  The square-root sampling takes the square root of class frequency to alleviate the imbalanced distribution, the probability p_j of class j is calculated as below

$$p_j = \frac{n_j^{\frac{1}{2}}}{\sum_{i=1}^{C} n_i^{\frac{1}{2}}} \tag{1}$$

  where n_j and n_i are class frequency for class j and i respectively, C is the number of classes.

Re-weighting methods:

- Class-balanced Loss [7]

  The class-balanced loss takes inverse effective number of samples as weights, it can be applied to existing loss functions such as cross-entropy (CE), binary cross-entropy (BCE) and focal loss *etc*.

  In our implementation, we use the class-balanced binary cross-entropy loss (CB_BCE) and set $\beta = 0.9999$.

$$L_{CB\_BCE}(\boldsymbol{z}, y) = \frac{1-\beta}{1-\beta^{n_y}} \sum_{i=1}^{C} \log(\frac{1}{1+exp(-z_i^t)}) \tag{2}$$

$$z_i^t = \begin{cases} z_i, & \text{if i=y.} \\ -z_i, & \text{otherwise.} \end{cases} \tag{3}$$

  where n_y is the number of samples in class y.

- LDAM Loss [3]

  LDAM loss encourages the network to have a class-dependent margin for multiple classes as

$$\gamma_j = \frac{C}{n_j^{1/4}} \tag{4}$$

  where C is a scaling constant, n_j is the number of samples for class j. We implement LDAM loss with DRW training schedule which defers the effect of re-weighting after 50 epochs, the largest enforced margin is set to 0.5.

- EQL [20]

  EQL ignores the gradient from head classes samples for tail classes, so that the network training would take equal importance for each class. We utilizes the sigmoid cross-entropy loss version of EQL formulated as

$$L_{EQL} = -\sum_{j=1}^{C} \omega_j \log(\hat{p_j}) \tag{5}$$

$$\omega_j = 1 - \beta T_\lambda(f_j)(1-y_j) \tag{6}$$

  where T_λ is a threshold function which is set to 1 for tail classes and 0 otherwise. β is a random variable and it is set to 1 with a probability of γ. We set $\gamma = 1.76 \times 10^{-3}$ for experiments.

Mixup: Mixup [22] is a popular data augmentation method that linearly interpolates two samples at pixel level and their targets. The formula is

$$\begin{aligned} \hat{x} &= \lambda x_i + (1-\lambda) x_j \\ \hat{y} &= \lambda y_i + (1-\lambda) y_j \end{aligned} \tag{7}$$

where (x_i, y_i) and (x_j, y_j) are two randomly sampled examples, $(\hat{x}, \hat{y})$ is the mixed input, $\lambda \sim Beta(\alpha, \alpha)$. We set $\alpha = 1$ for all experiments. Note that Mixup is performed on input space and the models are trained from scratch in recent literature [8, 21]. In contrast, we implement Mixup on 2D features. It would be interesting to implement Mixup on raw videos but it is out of the setting in this paper.

Two-stage methods: The two-stage methods aims to rectify the decision boundaries among head and tail classes, one simple yet effective way is to rebalance the weights of classifier. Following [10], we conduct experiments on three two-stage methods:

- τ-normalization

  It rectifies the decision boundaries by directly adjusting the norms of classifier weights as following:

$$\widetilde{\omega_i} = \frac{\omega_i}{\| \omega_i \|^\tau} \tag{8}$$

  where $\omega_i \in \mathbb{R}^d$ denotes the classifier weights of class i, $\widetilde{\omega_i}$ are the scaled weights, τ is the "temperature" of the normalization, we set τ to 1 in experiments.

- cRT (Classifier Re-training)

  The approach fixes the representations, randomly re-initialize the classifier weights and only re-trains the classifier with class-balanced sampling.

- LWS (Learnable Weight Scaling)

  A variety of τ-normalization is called LWS, it fixes both the representations and classifier weights and optimizes a scaling factor f_i, which can be written as

$$\widetilde{\omega_i} = f_i * \omega_i, \text{where } f_i = \frac{1}{\| \omega_i \|^\tau}$$

6. Extensions of Ablation Study

Clip length. We tested the influence of clip length to show our approach is generic. Experiments on different clip length show that clip length is not essential to our method. See Table 3.

clip length	baseline	FrameStack
60	0.499	0.516
120	0.497	0.515

Table 3. Influence of clip length (Tested on ResNet-50 backbone and Nonlinear model, overall mAP is reported).

7. Definition of Head, Medium and Tail Classes

We follow the general definition of head, medium and tail classes in [13]. We first count the number of videos for each class in the training set (because we usually can not obtain ground truth for test data in real-world), then we define 47 head classes ($\#videos > 500$), 617 medium classes ($100 < \#videos <= 500$) and 340 tail classes ($\#videos <= 100$). Details of head, medium and tail classes are provided in table 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15.

8. Full Taxonomy of VideoLT Dataset

VideoLT contains 256,218 untrimmed videos, annotated into 1,004 classes with a long-tailed distribution. There are 13 top-level entities including: *Animal*, *Art*, *Beauty and Fashion*, *Cooking*, *DIY*, *Education and Tech*, *Everyday life*, *Housework*, *Leisure and Tricks*, *Music*, *Nature*, *Sports* and *Travel*. The brief taxonomy of VideoLT is shown in figure 1. The full taxonomy system is shown in figure 2, 3, 4, 5, 6, 7, 8. VideoLT dataset will be made public soon.

HEAD CLASS	INDEX	#VIDEOS	HEAD CLASS	INDEX	#VIDEOS	HEAD CLASS	INDEX	#VIDEOS
armWrestling	58	724	eyeMakeup	311	569	sportStacking	853	574
backstroke	72	671	fireworksShow	334	563	sportsTrack	854	510
basketball	85	648	guitarPerformance	376	597	sprayPainting	855	544
batonTwirling	89	516	handstand	387	559	toyFigures	907	635
beach	90	1228	horse	406	683	violinPerformance	942	675
beatbox	97	597	hulaDance	415	673	vocalTraining	945	549
boating	115	536	lake	475	1098	windTurbines	984	537
bridge	124	572	laptop	477	546			
brushingTeeth	126	561	makingNecklaces	575	506			
bungeeJumping	140	573	mountain	639	1122			
cableCar	145	622	penSpinning	682	523			
camel	148	1221	river	753	1912			
camelRiding	149	612	ropeClimbing	765	551			
celloPerformance	177	682	sandboarding	774	552			
cookingBacon	229	621	singingOnStage	818	528			
cruiseShip	246	504	skateboarding	821	567			
desert	262	1268	smartphone	831	1002			
dog	276	1044	snowWeather	839	1121			
drumPerformance	289	698	soccer	843	630			
elephant	304	1043	softball	844	562			

Table 4. Table of 47 head classes ($\#videos > 500$)

MEDIUM CLASS	INDEX	#VIDEOS	MEDIUM CLASS	INDEX	#VIDEOS
3DPainting	0	137	bambooForest	78	101
3DPrinter	1	308	bambooRafting	79	291
AngkorWat	4	233	bananaRiding	80	239
BabyLearningToEatWithSpoon	5	212	banjoPerformance	81	303
BigBen	6	102	barbellWorkout	83	474
ChineseBrushWriting	7	117	baseball	84	209
ChineseFlutePerformance	9	110	bassPerformance	86	353
Colosseum	11	101	bat	87	117
EmergencyAmbulance	13	158	beachFootball	91	295
FrenchHornPerformance	15	140	beachVolleyball	93	151
GuineaPig	18	448	beachWalking	94	117
HandBellPerformance	19	243	bearingTheStretcher	96	130
IndianDance	20	384	beaver	98	180
KolnerDom	21	116	bee	99	269
LatinDance	22	451	bellyDance	102	341
RioDeJesus	27	155	bikeTricks	103	254
ScotlandDance	28	199	biking	104	207
SitarPerformance	29	131	billiard	105	269
TerracottaArmy	32	101	birthday	106	208
TowerBridge	34	168	blackCloud	107	290
TowerOfLondon	35	119	bloodGlucoseMeter	108	123
UAV	36	353	blowingASoapBubble	110	102
XylophonePerformance	38	199	blowingBubblegumBubbles	111	322
accordionPerformance	39	417	blowingUpABalloon	114	130
acrobatics	40	255	bodyWeightSquats	116	285
acropolisOfAthens	41	154	bowling	120	220
aircraft	43	189	boxing	121	278
airwheel	44	177	breaststroke	122	211
alpaca	45	169	bridalShower	123	120
americanFootball	46	287	bridgeJumping	125	236
amusementPark	47	394	buildingAGreenhouse	130	257
antelope	48	193	buildingASandcastle	132	281
applyingAPhoneScreenProtecter	49	278	bumperCars	139	485
applyingAWig	50	335	butterfly	141	242
applyingFacialMask	52	258	butterflyKnifeTricks	142	317
archOfTriumph	54	103	butterflyStroke	143	286
archery	56	278	caberTossing	144	254
arrangingFlowers	59	442	calmingACryingBaby	147	122
assemblingABike	62	225	canyon	153	312
assemblingAComputer	63	243	capybara	155	348
babyCrawling	67	241	carAccidents	156	415
babyEatingSnack	68	261	carExhibition	157	193
babyShower	69	139	carRacing	159	160
babySuckingFinger	70	106	carWashing	161	143
backMassage	71	168	cardManipulation	162	400
badminton	73	497	cardiopulmonaryResuscitation	164	176
bagpipePerformance	74	188	carriageDriving	165	469
bakingHam	75	139	carvingAPumpkin	166	344
balanceBeam	76	456	cat	167	452
ballet	77	414	cathedralExterior	174	209

Table 5. Table of 617 medium classes ($100 < \#videos <= 500$)

MEDIUM CLASS	INDEX	#VIDEOS	MEDIUM CLASS	INDEX	#VIDEOS
celebratingNewYear	176	134	decoratingChristmasTree	258	349
chamberMusic	178	228	deer	259	150
changingBrakePads	180	215	deliciousFood	260	275
changingGuitarStrings	181	143	dholPerformance	264	108
changingOil	182	175	diningAtRestaurant	267	395
changingTires	183	161	discusThrow	271	225
charging	184	173	diving	273	387
cheerleading	185	301	djingPerformance	274	141
chicken	186	118	dogGoingUpTheStairs	277	104
chinchilla	188	382	doingChemistryExperiments	279	385
choppingWood	189	248	doingGraffiti	280	226
chorus	190	450	doingTheLaundry	281	132
christmasParty	191	106	dollCollectings	282	135
classroom	195	183	dolphin	283	157
cleaningABathroom	197	288	donatingBlood	284	164
cleaningALaptopFan	200	140	drinkingBeer	287	156
cleaningAStove	205	176	driving	288	245
cleaningAToilet	206	214	dumbbellWorkout	292	148
cleaningCarpet	209	151	duneBashing	293	220
cleaningFloor	211	140	dyeingHair	294	101
cleaningFruitsAndVegetables	212	111	earthquake	295	232
cleaningJewelry	214	197	eatingPopcorn	297	103
cleaningOilPaintBrushes	215	135	eatingWithChopsticks	298	157
cleaningWindows	218	222	ebookReader	299	180
clickerTraining	219	111	egyptianPyramids	300	137
coach	222	170	eiffelTower	301	285
cockFighting	223	303	electricToothbrush	302	113
coinCollecting	224	101	electronicOrganPerformance	303	145
combingHair	225	163	elephantRides	305	259
communicatingWithSignLanguage	226	105	embroidery	306	292
congaDance	227	180	escalatorRide	308	104
cookingAnOmelette	228	142	exploding	309	135
cookingBeefSteak	230	321	eyeExercises	310	130
cookingLemonChicken	233	110	eyebrowThreading	313	137
cookingLotusRoot	234	103	faceMassage	315	226
cookingMashedPotato	235	224	facePainting	316	244
cookingPasta	237	252	fanDance	320	131
cookingRiceNoodles	239	120	fanMeeting	321	327
cookingTofu	241	187	fancyDressParty	322	102
cow	242	408	farewellParty	323	106
crab	243	182	fencing	325	373
crayonDrawing	244	313	ferret	326	424
crushingGarlic	247	237	fingerPainting	328	225
crying	248	201	fireBalloon	329	113
curingMeat	250	243	fireDance	330	271
cuttingFruits	251	305	fireFighting	332	118
danceDrama	254	145	firstAidForChocking	335	125
darts	255	155	fishing	336	355
debate	256	241	flutePerformance	343	242
debrisFlow	257	157	flyingKites	344	321

Table 6. Table of 617 medium classes ($100 < \#videos <= 500$)

MEDIUM CLASS	INDEX	#VIDEOS	MEDIUM CLASS	INDEX	#VIDEOS
foldingANapkin	346	296	infraredThermometer	431	138
foldingClothes	347	241	inkAndWashPainting	433	115
forest	348	160	insideAirplane	434	245
framingADoor	350	165	insideBus	435	187
frog	351	136	insideTheOrientalPearlTVTower	436	185
frontCrawl	352	130	installingASink	442	146
fruitTreePruning	353	270	installingCurtainRods	444	108
fryingEggplants	354	453	jetski	449	465
fryingFish	355	336	jobFair	450	180
fryingMushroom	356	323	judo	454	140
geocaching	361	190	jugglingBalls	455	179
gibbon	362	126	jumpingJack	456	151
giraffe	363	286	kangaroo	457	128
glacier	365	171	karstCave	458	123
goat	366	214	kayaking	459	198
golfing	369	234	kendo	460	247
gorilla	371	391	kickingShuttlecock	461	355
graduation	372	186	kidsFighting	462	135
growingCrystals	375	201	kidsMakingFaces	464	313
gymBall	377	409	kidsPlayingWithBlocks	467	391
hairCutting	380	169	kindergarten	468	198
hairstyleDesign	381	437	kiteSurfing	470	224
hammerThrow	383	190	knitting	472	346
hammering	384	123	lacingShoes	474	184
hamster	385	157	landsailing	476	413
handball	386	136	laserProjectionKeyboard	478	106
harmonicaPerformance	390	367	legMassage	480	245
harpPerformance	391	290	lemur	481	124
havingAcupuncture	392	130	leopard	482	125
headMassage	395	173	lightning	483	231
hearingAid	396	334	lion	484	255
highJump	399	101	lionDancing	485	207
hiking	400	289	longJump	486	121
hipHopDance	401	325	louvre	488	121
hippo	402	121	lunarEclipse	490	151
horizontalBar	405	369	lunges	491	370
horseRacing	407	109	makingABottleRocket	495	134
horseRiding	408	280	makingABrooch	497	127
hotel	411	155	makingACalendar	498	189
housePlants	412	119	makingACandleHolder	499	220
hugging	414	150	makingAChair	500	159
hulaHoop	416	479	makingACrown	501	279
hunting	418	181	makingAGiftBasket	503	102
hurdling	419	481	makingAHeadboard	504	101
iceBoating	420	367	makingALamp	506	106
iceBucketChallenge	421	444	makingAPillow	512	262
iceDancing	423	179	makingAPlanter	513	106
iceFishing	424	343	makingAShelf	516	103
iceHockey	425	286	makingASkirt	517	161
iceSkating	427	242	makingAWreath	518	221

Table 7. Table of 617 medium classes ($100 < \#videos <= 500$)

MEDIUM CLASS	INDEX	#VIDEOS	MEDIUM CLASS	INDEX	#VIDEOS
makingApplePie	519	164	makingPotatoCakes	587	124
makingBananaChips	520	148	makingRings	589	163
makingBasket	521	117	makingRug	590	167
makingBathBombs	522	279	makingSalad	591	217
makingBoiledEggs	524	130	makingSandwich	592	258
makingBookmark	525	301	makingSashimi	593	250
makingBouquets	526	104	makingScrambledEggs	595	361
makingBracelets	527	316	makingShorts	597	242
makingBread	528	296	makingSkinCream	599	227
makingBurritos	529	122	makingSnowman	600	277
makingButter	530	135	makingSoap	601	214
makingCake	531	257	makingSorbet	602	272
makingCandle	532	180	makingSoup	603	417
makingCandyApples	533	237	makingSteamedStuffedBun	607	112
makingCeramicCraft	535	496	makingSushi	609	254
makingChineseDumplings	537	217	makingTacos	610	155
makingCider	539	157	makingTea	611	167
makingCoffee	540	258	makingTheBed	612	139
makingCongee	541	159	makingToast	613	436
makingCookies	542	255	makingViolin	614	108
makingCrepes	543	274	makingWaffles	615	329
makingCurryChicken	544	148	makingWallet	616	218
makingDonuts	547	223	makingYogurt	617	134
makingEarrings	548	347	manatee	618	174
makingFestivalCards	550	181	mantaRay	619	212
makingFrenchFries	552	141	mantis	620	111
makingFriedChicken	553	133	mantisShrimp	621	345
makingFriedEggs	554	157	marathon	623	216
makingFriedRice	555	322	marchingBand	624	374
makingHairPins	557	125	marriageProposal	625	170
makingHalloweenCustumes	558	117	martial	626	153
makingHamburgers	559	322	massageChair	627	267
makingHotChocolate	560	385	mattressJumping	628	138
makingIceCream	563	154	medalPresentation	629	135
makingJuice	564	293	militaryParade	632	367
makingLemonade	567	116	motocross	638	140
makingMilkTea	569	230	mountingATV	640	280
makingMilkshakes	570	393	mowing	641	208
makingMixedDrinks	571	172	museum	642	239
makingMuttonCurry	574	272	nailArtDesign	643	326
makingNoodlesFromScratch	576	141	nebulizer	644	146
makingPancakes	577	261	nightMarket	645	187
makingPaperFlowers	578	152	oilPainting	648	398
makingPaperPlane	579	398	openingABottle	649	125
makingPencilCases	580	179	operatingAChainsaw	653	108
makingPhoneCases	581	194	oralQuiz	655	148
makingPhotoFrame	582	274	ostrich	659	172
makingPickles	583	166	outsideAirplane	660	239
makingPizza	584	178	outsideBus	661	145
makingPlasticineModel	585	247	outsideTheOrientalPearlTVTower	662	124

Table 8. Table of 617 medium classes ($100 < \#videos <= 500$)

MEDIUM CLASS	INDEX	#VIDEOS	MEDIUM CLASS	INDEX	#VIDEOS
panda	668	164	rainbow	739	103
pantomime	670	252	rapMusic	740	366
parade	672	265	rehabilitation	741	175
paragliding	673	105	repairingAMobile	744	138
parallelBars	674	289	repairingARefrigerator	745	246
parkingCars	675	198	repairingAWasher	746	212
parkour	676	302	repairingMusicalInstruments	748	200
parrot	677	221	repalcingInkCartridges	749	111
peacock	678	112	residentialWiring	750	130
peelingOnion	679	189	ridingStationaryBicycle	752	106
penCalligraphy	680	275	roastingCorn	755	335
pencilPainting	683	300	roastingFish	756	217
penguin	684	104	roastingPeppers	757	260
photography	685	169	roastingTurkey	758	119
pianoPerformance	687	402	rockBandPerformance	759	359
pickingCorns	688	174	rockClimbing	760	125
pig	691	212	rollerHockey	763	342
pigRacing	692	146	rollerSkating	764	205
pigeonRace	693	110	ropeSkipping	766	331
pipaPerformance	694	176	ropesCourse	767	327
pitchingATent	695	252	rowing	768	414
plantingATree	696	128	safes	770	133
plastering	697	238	salsaDance	771	473
playABalloonPrank	699	199	sandArt	772	394
playground	701	231	sandStorm	773	111
playingChess	702	139	saxophonePerformance	775	294
playingFrisbeeWithDog	704	231	screenPrintting	776	101
playingFrisbeeWithPeople	705	251	scubaDiving	777	202
playingPuzzle	707	101	sculpting	778	103
playingRacquetball	708	146	seaLion	780	193
playingWaterGun	709	176	settingTheTable	782	288
playingWithNunChucks	710	316	sewingAButton	783	177
playingWithRemoteControlledAircraft	711	264	shadowPlay	784	147
playingWithRemoteControlledCars	712	146	sharpeningAPencil	786	141
poker	714	143	sharpeningKnives	787	242
polarBear	715	144	shavingBeard	788	134
poleVault	717	149	shearingASheep	790	204
polishingShoes	718	125	sheep	791	415
pommelHorse	720	329	sheepHerding	792	136
poultryFarming	721	137	shipwrecks	794	243
printmaking	724	129	shooting	795	407
publicSpeech	726	197	shovelingSnow	797	143
pullUps	727	471	showingFashionableEarrings	799	167
punchingBagWorkout	728	211	showingFashionableHandbags	802	209
puppetShow	729	237	showingFashionableHighHeels	804	233
pushUps	730	155	showingFashionableScarfs	809	204
puttingOnShoes	734	246	shuffleboard	814	106
rabbit	735	374	singingInKtv	816	308
raccoon	736	226	singleLensReflexCamera	819	172
rafting	737	295	sitUps	820	242

Table 9. Table of 617 medium classes ($100 < \#videos <= 500$)

MEDIUM CLASS	INDEX	#VIDEOS	MEDIUM CLASS	INDEX	#VIDEOS
skiing	822	284	theLeaningTowerOfPisa	897	232
skydiving	823	169	thePalaceOfVersailles	898	164
sledding	824	252	theStatueOfLiberty	899	278
sleepoverParty	825	104	thumbWrestling	900	120
slippingOnIce	826	296	tiger	901	188
smartBracelet	827	300	tireWorkout	902	138
smartTV	830	118	tornado	904	136
smokeTricks	833	133	townHallMeeting	905	112
snail	834	227	toy	906	105
snake	835	473	traditionalOpera	908	137
snowRafting	837	153	trafficJam	909	122
snowballFight	840	392	trafficPoliceman	910	180
solarEclipse	845	119	train	911	417
solvingMagicCube	846	281	trampolineJumping	912	236
soccerJuggling	847	133	treadmill	915	244
sportAerobics	852	440	treeClimbing	916	196
squareDance	857	151	tricycleRiding	919	204
squirrel	858	305	trimmingGardenPlants	920	221
stampCollecting	859	251	trombonePerformance	921	445
starryNightSky	860	125	trumpetPerformance	922	364
stealing	861	140	tsunamiWave	923	179
stillRings	863	315	tugOfWar	924	174
streetFighting	864	155	turtle	925	152
stringArt	865	222	tyingABow	926	101
submarine	867	355	tyingATie	927	206
sumoWrestling	868	399	ukulelePerformance	930	263
sunset	869	431	umbrellaDance	931	123
suonaPerformance	870	248	undergroundSubway	932	344
surfing	871	155	unevenBars	933	308
swearingAnOath	873	130	unicycleTricks	934	439
sweepingRobot	874	137	useAWheelchair	936	106
swordDance	875	107	usingAFireSteel	937	196
symphonyOrchestraPerformance	876	175	vehiclesCrossingRivers	940	226
synchronizedSwimming	877	402	velometer	941	144
tablaPerformance	878	273	volcanoEruption	946	142
tableTennis	879	355	volunteerWork	948	238
tabletPC	880	441	walkingWithDog	951	116
tackingUpAHorse	881	126	wallPainting	952	133
taekwondo	882	419	wallPushup	953	124
taiChiChuan	883	458	waltzDance	955	292
tailgateParty	884	230	warmUpActivities	956	275
tajMahal	885	221	washingAnInfant	957	162
takingMakeupOff	886	115	washingDishes	958	160
tango	887	412	washingShoes	963	169
tattooing	888	439	waterPolo	965	458
taxi	889	271	watercolourPainting	966	403
temple	893	224	waterfall	967	218
tennis	894	186	wearLipstick	969	468
tetherball	895	154	weddingCeremony	970	157
theGreatWall	896	177	weddingDance	971	209

Table 10. Table of 617 medium classes ($100 < \#videos <= 500$)

MEDIUM CLASS	INDEX	#VIDEOS
weddingPhotography	972	147
weddingReception	973	156
wetland	975	272
wheelchairBasketball	976	158
wheelchairFencing	977	144
wheelchairRace	978	144
windSailing	983	452
wineryTour	986	279
wingsuitFlying	987	148
woodCarving	992	127
woodpecker	993	146
wrappingAWound	994	184
wrappingTheGift	995	440
xinjiangDance	997	334
yoga	999	294
yoyoTricks	1000	404
zumbaDance	1003	236

Table 11. Table of 617 medium classes ($100 < \#videos <= 500$)

TAIL CLASS	INDEX	#VIDEOS	TAIL CLASS	INDEX	#VIDEOS
ACappella	2	87	camping	151	85
ATM	3	60	canteen	152	53
ChineseChess	8	60	captureTheFlag	154	79
ChineseJumpRope	10	89	carPainting	158	59
DafPerformance	12	80	carTowing	160	61
EmpireStateBuilding	14	56	cardingWool	163	58
GPSNavigationDevice	16	62	cat'sCradle	168	50
GreatGoldenStupa	17	47	catTakingABath	169	99
LeshanGiantBuddha	23	57	catUsingLitterBox	170	76
MountRushmoreNationalMemorial	24	70	catchingACrab	171	69
PSP	25	99	catchingAFrogInAPond	172	68
PotalaPalace	26	95	catchingASquirrel	173	83
Stonehenge	30	72	cattleHerding	175	59
SydneyOperaHouse	31	100	changingABabyDiaper	179	100
TheFobiddenCity	33	53	chimePerformance	187	58
XunPerformance	37	68	circusAnimalPerformance	192	95
airCrash	42	81	clappingHands	193	81
applyingConsealer	51	67	clarinetPerformance	194	86
applyingPesticides	53	60	clawMachineTricks	196	60
archaeologist	55	94	cleaningABicylce	198	98
arcticFox	57	83	cleaningAChimney	199	75
artificalLimb	60	62	cleaningAMicrowaveOven	201	89
ass	61	97	cleaningAMonitorScreen	202	63
assemblingACrib	64	89	cleaningAMotorcycle	203	75
astronaut	65	73	cleaningASilverware	204	72
auroraPolaris	66	91	cleaningAWhiteboard	207	58
barbecue	82	59	cleaningAnAirConditioner	208	67
bathingDog	88	69	cleaningCurtains	210	100
beachSwimming	92	88	cleaningGlasses	213	87
bear	95	91	cleaningOutClosets	216	87
beekeeping	100	94	cleaningPaintedWalls	217	57
beggingForMoney	101	77	cliffDiving	220	61
blowingAConchShell	109	81	clock	221	66
blowingDryHair	112	88	cookingFriedPorkChop	231	81
blowingFire	113	58	cookingGooseLiver	232	94
bodyguard	117	57	cookingOffal	236	94
boilingFish	118	80	cookingPorkLegs	238	88
bookshop	119	53	cookingSeafood	240	96
buildingACabinet	127	74	cricket	245	66
buildingAFence	128	63	cudgel	249	57
buildingAFirePit	129	46	cuttingNail	252	64
buildingAPoleBarn	131	74	cymbalPerformance	253	87
buildingAShed	133	55	departureHall	261	59
buildingATable	134	87	detective	263	77
buildingAnOutdoorBench	135	84	diaboloTricks	265	74
buildingBrickColumns	136	56	diggingOutACarFromSnow	266	64
buildingConstruction	137	96	dinnerAtHome	268	86
bullFighting	138	60	disabledWaterSkiing	269	67
calligraphyExhibition	146	87	discGolf	270	68
camelWresting	150	59	dishwasher	272	64

Table 12. Table of 340 tail classes ($\#videos <= 100$)

TAIL CLASS	INDEX	#VIDEOS	TAIL CLASS	INDEX	#VIDEOS
dodgeball	275	71	iceClimbing	422	90
dogRacing	278	81	iceRain	426	98
dormitory	285	76	icebergCollapse	428	70
dove	286	79	imageScanner	429	58
duck	290	79	industrialRobot	430	82
duckTour	291	65	injectionForKids	432	95
earthworm	296	51	installAGarageDoor	437	66
energySavingBulbs	307	70	installCeramicTileFloor	438	81
eyebrowTatto	312	62	installFloorHeating	439	54
eyelashGrafting	314	80	installingABathtub	440	52
faceSteaming	317	66	installingACeiling	441	71
fallingTide	318	60	installingAWindow	443	58
familyNight	319	60	internetCafe	445	64
fasteningTheSeatBelt	324	84	ironingClothes	446	76
fertilizingAGarden	327	61	javelinThrow	447	98
fireExtinguisher	331	70	jellyfish	448	45
fireplace	333	65	jobInterview	451	70
fittingTiles	337	66	journalist	452	96
fixingAChair	338	89	jousting	453	93
fixingBrokenWindows	339	54	kidsGettingDressed	463	55
fixingLeaks	340	83	kidsPlayingDoctor	465	78
flagRaising	341	68	kidsPlayingOnTheSlide	466	88
flood	342	75	kissing	469	76
fog	345	98	knifeFighting	471	66
fox	349	83	koala	473	94
fryingRiceCake	357	70	lavenderGarden	479	67
fuelFilling	358	80	longSleeveDance	487	80
funeral	359	100	lumbering	489	78
gasing	360	73	machinist	492	76
givingGifts	364	56	maglevTrain	493	61
goldPanning	367	59	makingABookCover	494	70
goldfish	368	93	makingABrickWalkway	496	61
goose	370	82	makingADreamCatcher	502	61
graftingATree	373	66	makingAKeyHolder	505	58
groupBanquet	374	74	makingALeprechaunTrap	507	75
hackALighter	378	77	makingAMousepad	508	50
hackathon	379	58	makingANotebook	509	53
halloweenParty	382	65	makingAPaperPuppet	510	76
hangingAPictureOnTheWall	388	56	makingAPiggyBank	511	90
hangingWallpaper	389	79	makingAPotHolder	514	57
hawk	393	76	makingARubberStamp	515	68
head-mountedDisplay	394	70	makingBeer	523	54
hedgehog	397	85	makingCaramel	534	67
hide-and-seek	398	68	makingCheese	536	84
holdBreathChallenge	403	87	makingChocolate	538	73
homeTheatre	404	100	makingCurtains	545	64
hostingAShow	409	82	makingDolls	546	44
hotSpring	410	88	makingEggTarts	549	82
housewarmingParty	413	57	makingFire	551	85
humidifier	417	67	makingGloves	556	85

Table 13. Table of 340 tail classes ($\#videos <= 100$)

TAIL CLASS	INDEX	#VIDEOS	TAIL CLASS	INDEX	#VIDEOS
makingHotPot	561	56	polo	719	87
makingHotdog	562	86	pouringConcrete	722	59
makingKite	565	90	pressConference	723	90
makingLace	566	64	pub	725	57
makingMasks	568	57	puttingAirInTires	731	60
makingModelCars	572	73	puttingOnATeethRetainer	732	96
makingMooncakes	573	77	puttingOnContactLenses	733	75
makingPopcorn	586	69	rainWeather	738	56
makingRiceDumplings	588	56	removingDrywall	742	63
makingSausage	594	58	removingMold	743	57
makingShoes	596	100	repairingAWatch	747	61
makingSkewers	598	77	rhinoceros	751	79
makingSoybeanMilk	604	93	roadConstruction	754	64
makingSpringRolls	605	86	rockPainting	761	55
makingSteamedEgg	606	77	rocket	762	63
makingStrawberrySauce	608	57	rugby	769	77
maracasPerformance	622	60	seaHorse	779	94
metalDesign	630	57	seaStar	781	53
microscope	631	61	shark	785	59
militaryTraining	633	66	shavingLegs	789	80
milkingACow	634	93	shellColecting	793	61
milkingAGoat	635	58	shoppingMall	796	84
monkey	636	81	showingFashionableDownJackets	798	72
motionSensingGame	637	55	showingFashionableFullDresses	800	51
oboePerformance	646	78	showingFashionableGlasses	801	97
octopus	647	86	showingFashionableHats	803	54
openingACan	650	52	showingFashionableJackets	805	85
openingCelebration	651	85	showingFashionableJeans	806	84
opera	652	60	showingFashionableNecklaces	807	73
operatingRoom	654	76	showingFashionableRings	808	100
orderingFood	656	68	showingFashionableShirts	810	65
organizingCables	657	57	showingFashionableSkirts	811	53
organizingShoes	658	70	showingFashionableSuits	812	67
owl	663	62	showingFashionableWatches	813	84
packingBoxes	664	57	sidewalkSweeping	815	67
paintBalling	665	90	singingInStudio	817	61
palanquin	666	60	smartGlasses	828	94
panFlutePerformance	667	84	smartPen	829	99
pangolin	669	99	smokeDetector	832	78
paperCutting	671	91	snorkeling	836	82
penPainting	681	93	snowVolleyball	838	66
physicalExamination	686	84	snowplows	841	73
pickingFruits	689	55	soakFeet	842	84
picnic	690	92	spearplay	848	55
platypus	698	51	speedWalking	849	72
playOnTheSeasaw	700	54	spider	850	60
playingCraps	703	63	spinningBasketball	851	88
playingMahjong	706	87	spreadMulch	856	64
plough	713	90	stepDance	862	67
polaroid	716	77	stripPaint	866	73

Table 14. Table of 340 tail classes ($\#videos <= 100$)

TAIL CLASS	INDEX	#VIDEOS
swan	872	80
teaParty	890	53
teamBuildingActivities	891	90
telescope	892	57
toeWresting	903	63
trampolinedunk	913	88
transplantRiceSeedings	914	85
trialCourt	917	95
trianglePerformance	918	78
tyingRopeKnots	928	92
typing	929	66
usbFlashDrive	935	56
vaccinating	938	84
vacuumCleaner	939	69
visitingGrave	943	99
visitingPatients	944	58
volleyball	947	85
walkingOnStilts	949	84
walkingWithAStick	950	50
walrus	954	81
washingFace	959	77
washingFeet	960	67
washingHair	961	68
washingHands	962	65
waterFilledBallon	964	53
wateringALawn	968	59
welcomeParty	974	83
wheelchairSoftball	979	58
wheelchairTennis	980	82
whistle	981	52
whistlingWithFingers	982	76
windsorCastle	985	77
wirelessHeadphones	988	93
wirewalking	989	73
wolf	990	86
woodBurning	991	62
writingOnBoard	996	93
yangko	998	72
zebra	1001	98
zhengPerformance	1002	69

Table 15. Table of 340 tail classes ($\#videos <= 100$)

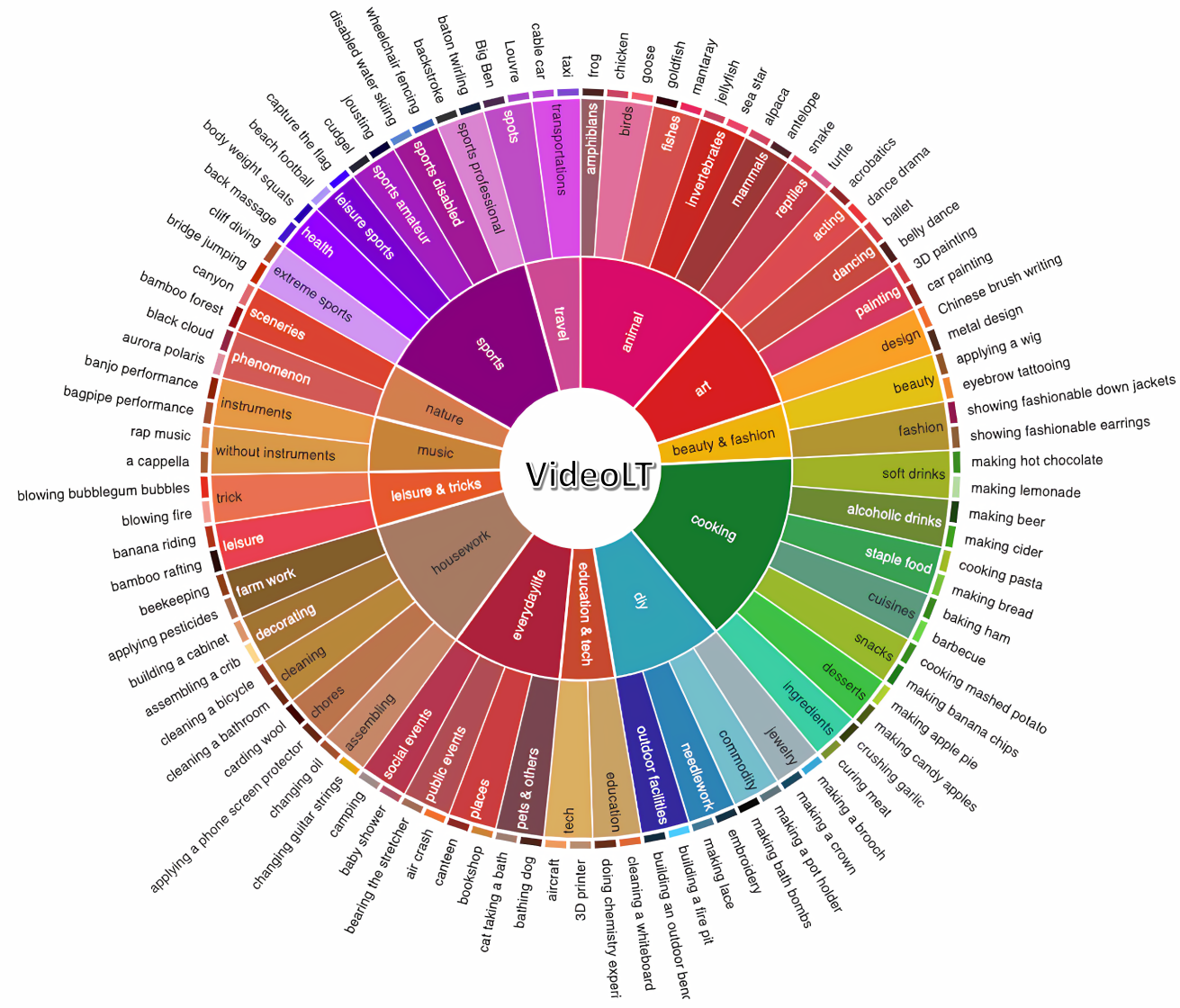

Figure 1. The taxonomy structure of VideoLT. There are 13 top-level entities and 48 sub-level entities, the children of sub-level entities are sampled.

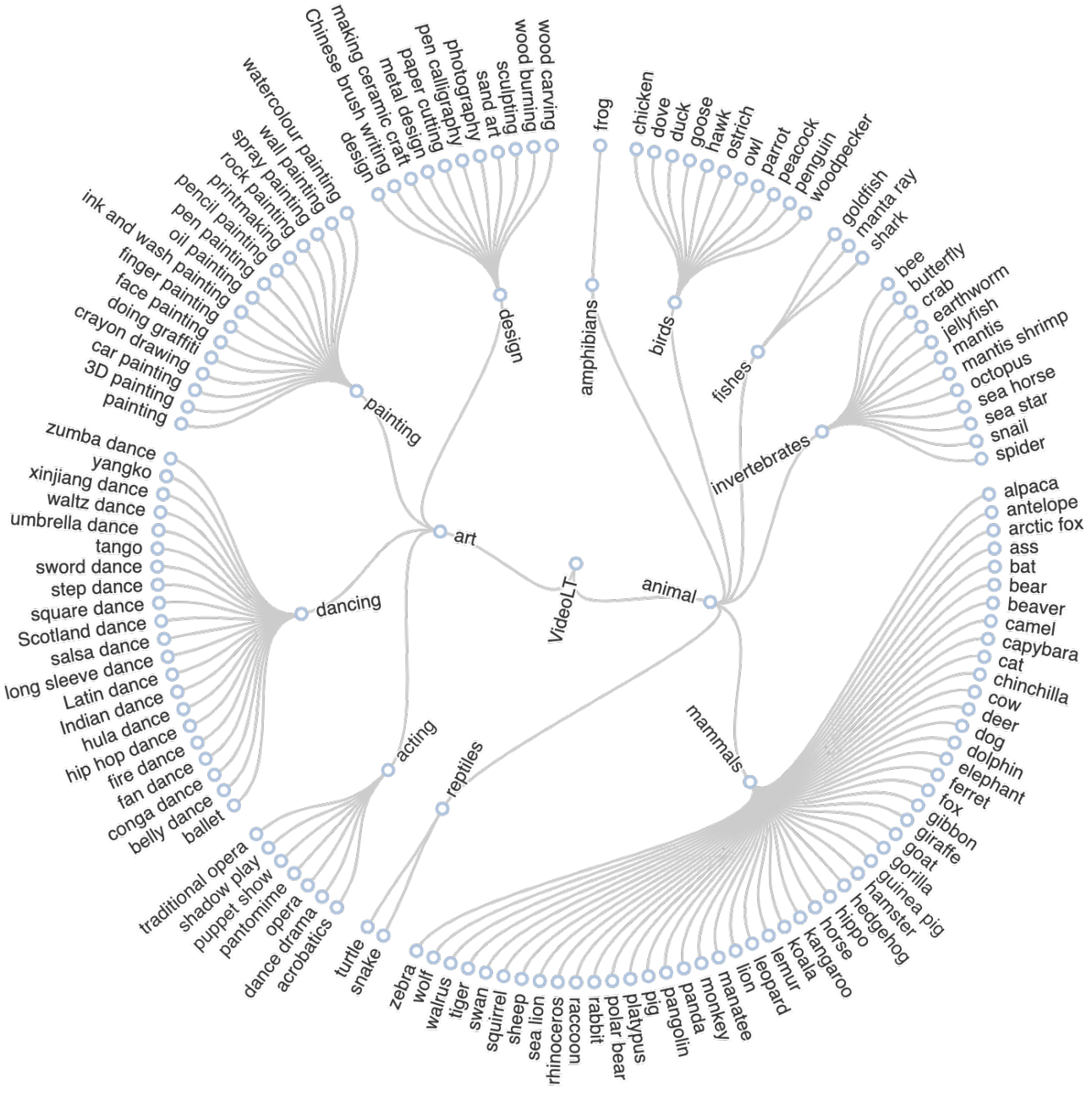

Figure 2. Taxonomy of top-level entities: *animal (#79)* and *art (#53)*

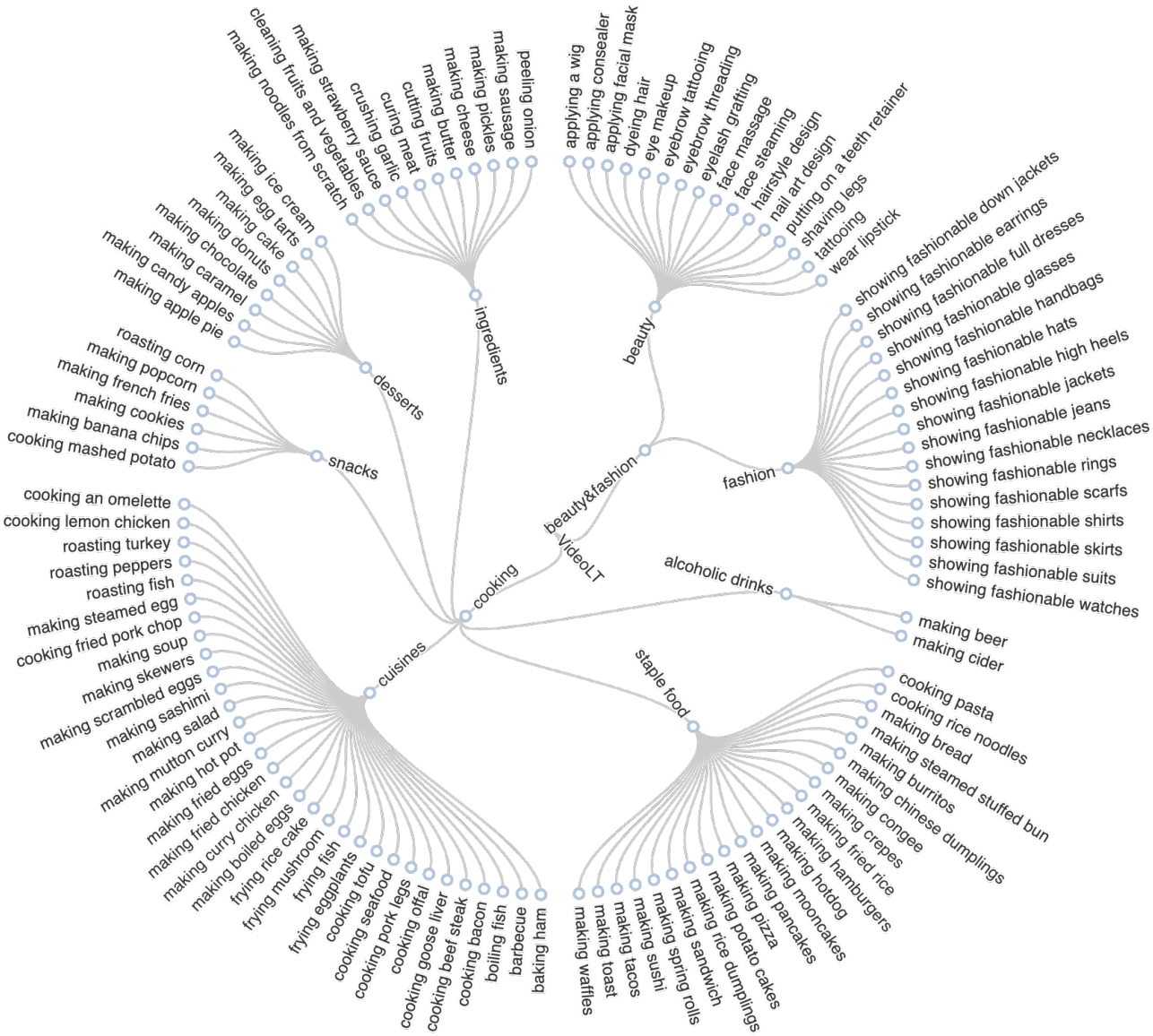

Figure 3. Taxonomy of top-level entities: *beauty&fashion (#32)* and *cooking (#92)*

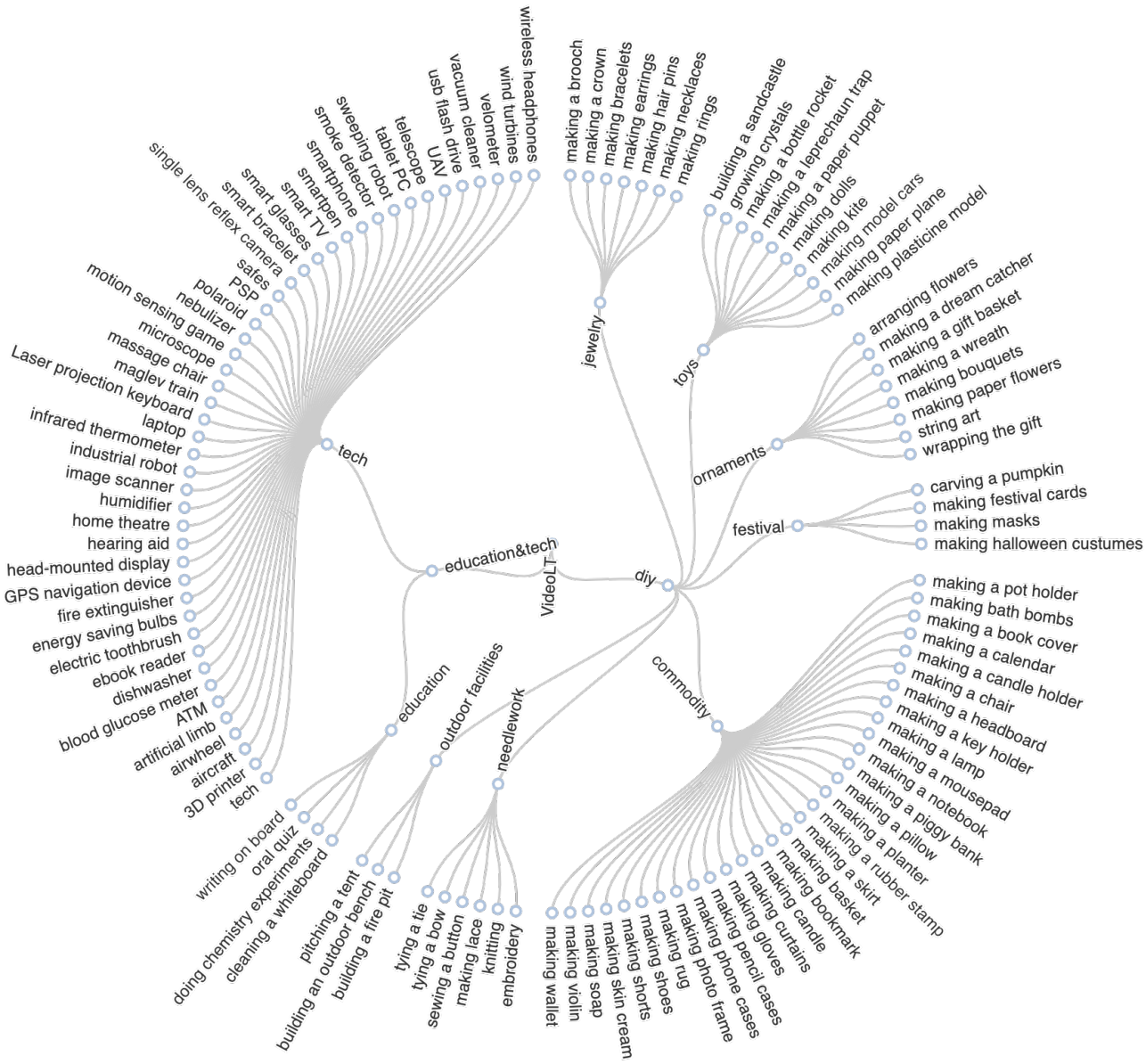

Figure 4. Taxonomy of top-level entities: *diy (#69)* and *education&tech (#49)*

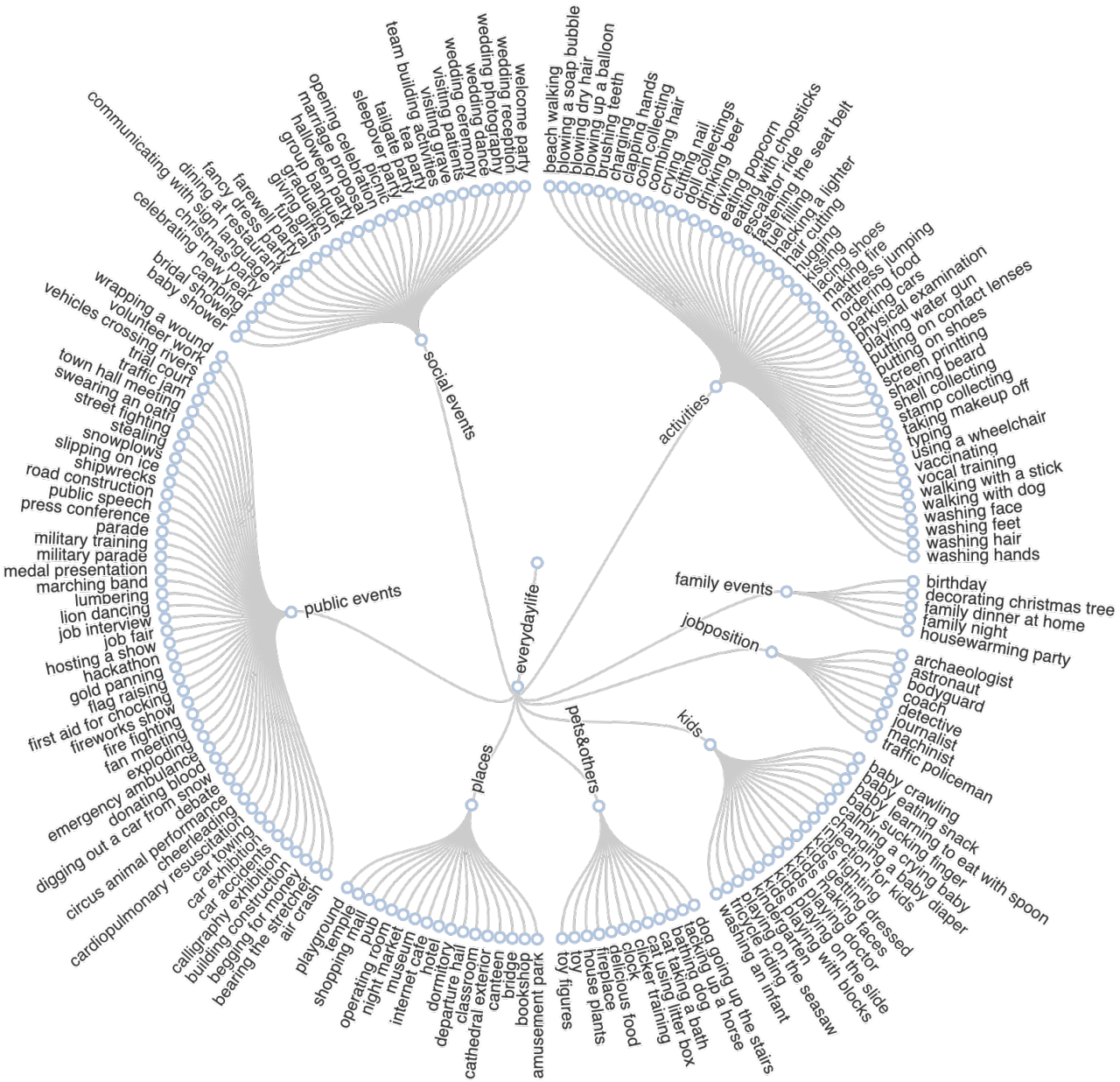

Figure 5. Taxonomy of top-level entities: *everydaylife (#182)*

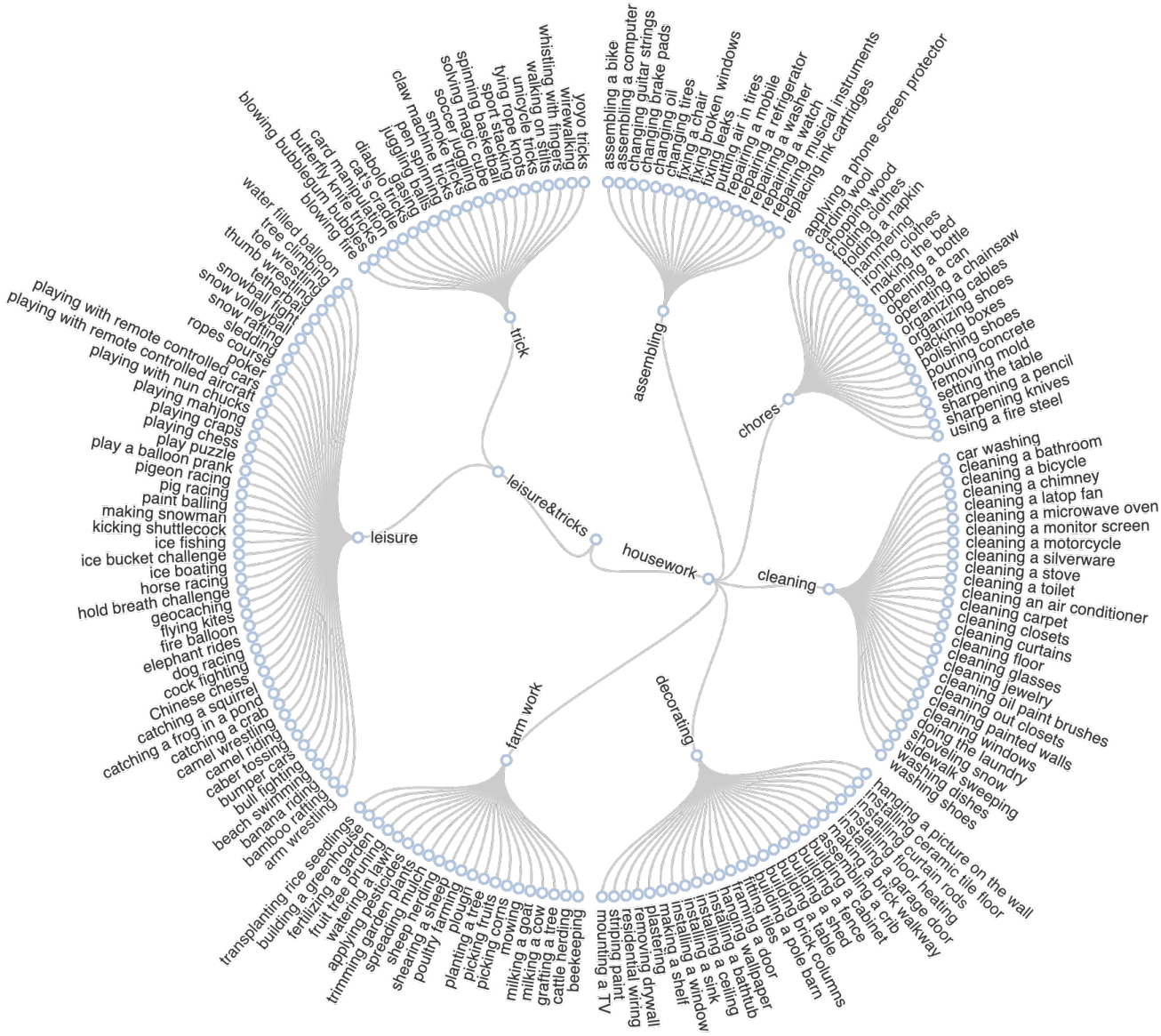

Figure 6. Taxonomy of top-level entities: *housework (#111)* and *leisuretricks (#69)*

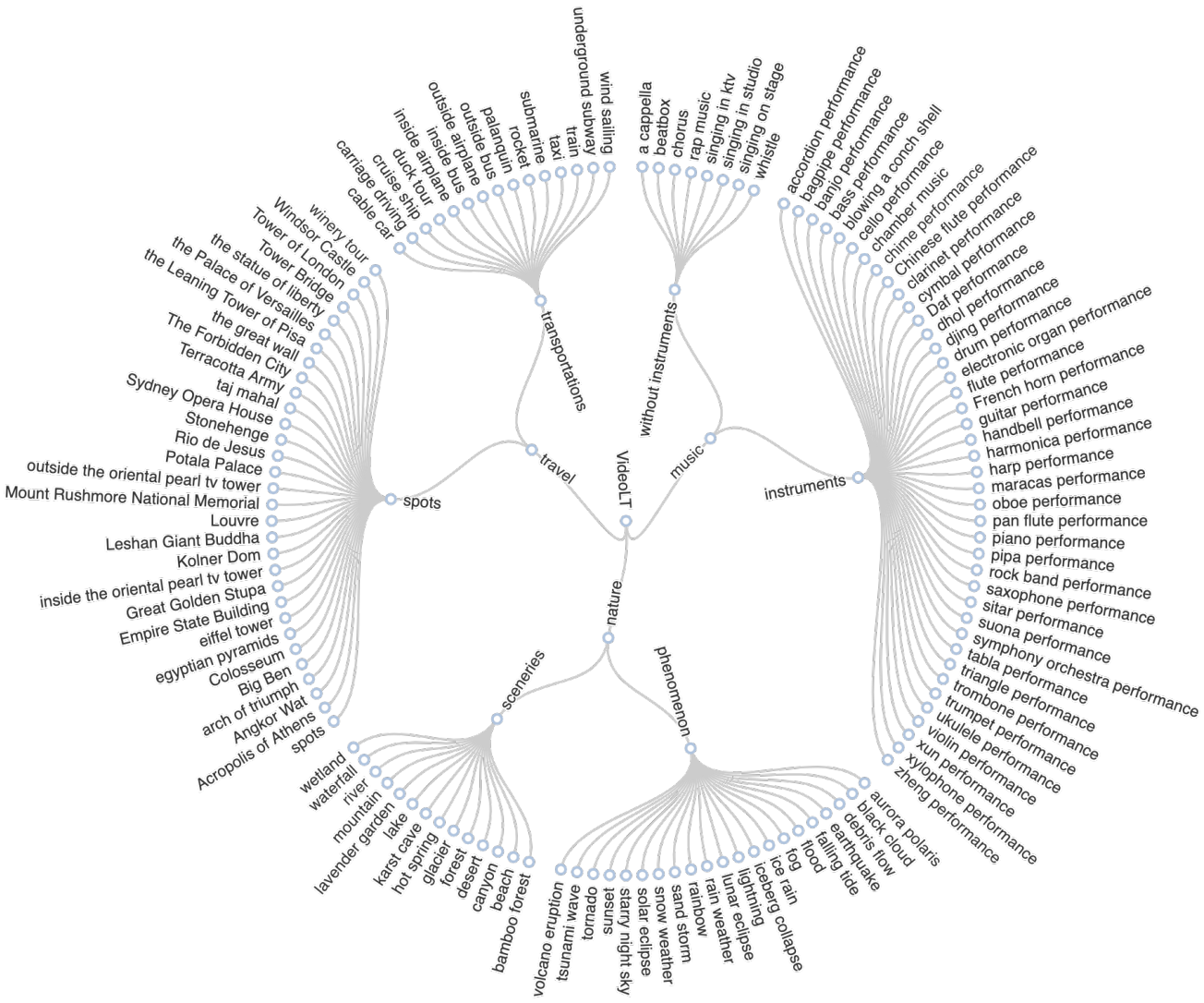

Figure 7. Taxonomy of top-level entities: *music (#49)*, *nature (#35)* and *travel (#45)*

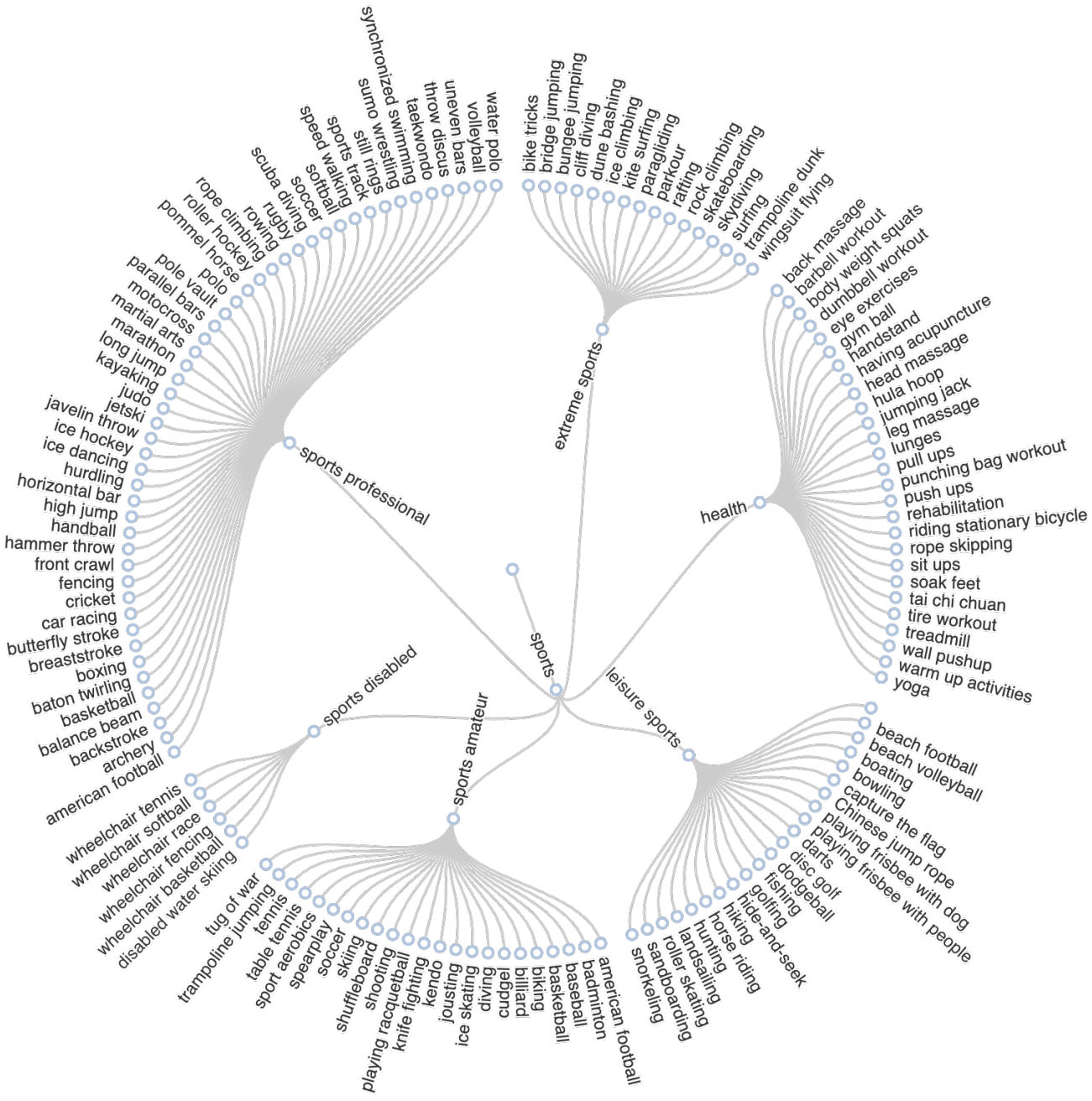

Figure 8. Taxonomy of top-level entities: *sports (#139)*

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015. 2

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *NIPS*, 2019. 3

[4] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 2

[5] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019. 2

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 2

[7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 3

[8] Haodong Duan, Yue Zhao, Yuanjun Xiong, Wentao Liu, and Dahua Lin. Omni-sourced webly-supervised learning for video recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 670–688. Springer, 2020. 3

[9] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, 2017. 2

[10] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020. 1, 2, 3

[11] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014. 2

[12] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, 2011. 2

[13] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 4

[14] Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, and Roland Memisevic. On the effectiveness of task granularity for transfer learning. *arXiv preprint arXiv:1804.09235*, 2018. 2

[15] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *CVPR*, 2009. 2

[16] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE TPAMI*, 2019. 2

[17] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *ECCV*, 2016. 2

[18] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, 2016. 2

[19] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[20] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *CVPR*, 2020. 3

[21] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, and Jinhyung Kim. Videomix: Rethinking data augmentation for video classification. *arXiv preprint arXiv:2012.03457*, 2020. 3

[22] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 3

[23] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *ICCV*, 2019. 2