Human Baseline for Zero-Shot Transfer Learning of Good Dogs

William Boag, Shashank Srikant
MIT CSAIL
Cambridge, MA, USA
{wboag,shash}@mit.edu

Abstract

It is estimated that there are 900 million dogs in the world. That is a lot of dogs, and we assure you that each one is a good dog [2]. Unfortunately, we don't have enough time to meet each dog de novo, and have inevitably needed to rely on word-of-mouth to learn about which dogs to meet and when. In this work, we demonstrate human-level performance for zero-shot dog recognition from features described by other humans. Human performance is robust (>85% accuracy), even when presented with challenging comparisons. This accuracy is in the same ballpark as Karpathy et al.'s work on a human baseline for ImageNet [6]. We believe that this work will help future researchers develop AI-based tools for super-human performance on word-of-mouth-based human-dog introductions. From a neuroscience perspective, this work also establishes the presence of a seeming information barrier between the visual cortex and the language system of the human brain.

Keywords Good Dogs, Zero-Shot Transfer Learning, Human Performance, Visual Cortex, Language Processing

Figure 1. Cover for the Zero-shot Dog Challenge slide deck

1 Introduction

There are so many dogs. It is a very exciting time to be alive, for sure. But with great opportunities come great opportunity costs: **the average citizen simply does not have enough time to meet every dog cold.** Typically, we get to know dogs through a third-party, usually starting with a verbal description. Many dogs are never photographed, which makes sense because very few dogs have sufficient Instagram followings. Dogs like Doug the Pug [3] are the exception rather than the rule. Most dogs are more like hidden gems, rather than national crown jewels. As such, it is critical that we are able to rely on verbal descriptions of dogs in order to learn which dog we will be meeting.

Everyone knows that technology-driven solutions will facilitate human-dog interactions, but it has not been rigorously studied how well humans perform on this task (and therefore what the value-add would be for AI-based technologies). In this work, we provide the much-needed empirical analysis of how well humans are able to identify dogs based on just verbal descriptions. We expect that with this groundwork laid, there will be an uptick in both research grants and startups for Computational Dog-related Studies. We speculate that this will likely be a hot area of research in the future.

Skeptics argue that these are problems of the past and that technology will enable picture-based social media networks for meeting dogs. In fact, there have been attempts of this in recent years such as *Tinder for Dogs* [10] and *Meet My Dog App* [8]. However, these technocrats overlook basic, fundamental flaws with this approach. For starters, some dogs are shy and don't want their photos online. In addition, no courts have ruled whether the Fourth Amendment (United States) or GDPR (European Union) applies to dogs. The United States legal system has recognized that some animals do have standing to bring suits in federal courts [4], so the matter of whether dogs have a legal right to privacy is currently unresolved. Nevertheless, such tools will not solve all of our problems.

Our contributions in this work are as follows

- We show that humans are pretty good at identifying dog breeds based on verbal comparative descriptions. These descriptions are produced by observing just 2-3 images of the dogs in question (and hence, zero-shot).
- By quantifying the performance of humans on a zero-shot transfer learning task, we establish a bound on what can possibly be performed by an AI technology designed for such tasks.
- We hypothesize that there exists a barrier between the information processed by the visual cortex, and the

SIGTBD, CSAIL, MIT, USA, 2019
2019.

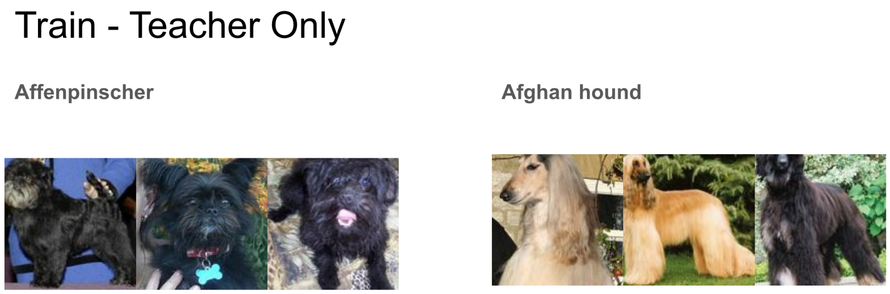


Figure 2. Example training slide shown to a teacher.

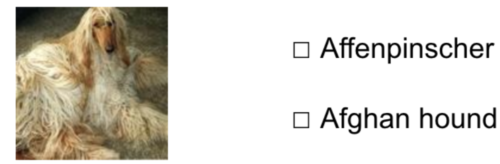


Figure 3. Example prediction slide shown to a team.

Figure 4. Task details. Two breeds of dogs, each with 3-5 images, were displayed to the *teacher* from each team. The rest of the team would not have access to these images while the *teacher* viewed this set of images. The *teacher* would then have to explain the breeds to the team. The teams would then be shown a test image and would be asked to guess the breed.

processed information being available to the language system of the human brain.

2 Previous Work

SIG TBD is home to seminal work in the field of Machine Learning for Dogs: Boag (2018) empirically proved that every dog in Cambridge, MA is a good dog [1], though it is still an open problem whether this has a theoretical basis to it.

In 2018, Kaggle unveiled a competition for Dog Breed Identification Challenge [5]. Similarly, Machine Learning techniques have been deployed for app-based breed classification [9]. Of course for all of these works, there is always an extensive training set to learn from hundreds of examples per class. This requires a lot of effort.

Dog-based Machine Learning has also arguably been studied in fields like Interpretable AI. Specifically, there are some works which used multiple pictures of dogs in their Computer Vision papers [11]. There may be other such papers as well.

More recently, attempts have been made to learn machine learning models through comparison [12]. This is closest in spirit to the human neural activity that the zero-shot transfer learning task in our work engages.

3 Methods

3.1 The Task

In February 2019, the authors participated in a competition among humans for zero-shot dog breed classification. The rules were as follows:

1. 5 rounds of game per team (each team comprised 6 people)
2. In each round, one person from a team will be a "teacher". The teacher is shown two dog breeds. For each breed, 2-5 images are shown. S/he is given 30 seconds to look at the two sets of images, and 30 seconds to describe the two breeds based on their observation.
3. The organizers then choose on unseen image of any of the two breeds and show it to the team members. Based on the descriptions they heard, the team members are now required to guess which of the two breeds the dog belongs to. They have a 50% probability of getting it right. They are then given 20 seconds to think/deliberate/discuss. Importantly, they have no access to the 'training set' which the teacher had access to.
4. All images are from Stanford Dogs Dataset [7]

An example of this procedure can be seen in Figures 2 (teacher training) and 3 (prediction slide). In this case, the teacher (while looking at Figure 2) might describe the difference between the dogs as "Affenpinschers are always black, whereas Afghan Hounds can sometimes be black and sometimes be brown" or perhaps "The Afghan Hound is a lot larger than the Affenpinscher." After the teacher describes the differences for 30 seconds, the team then sees Figure 3 and must deliberate to decide whether the dog is an Affenpinscher or Afghan Hound.

3.2 Teams

There were three teams that participated in this competition. Although all teams followed the same framework (i.e. take notes while teacher describes dogs and then reference those notes during prediction time), they employed different strategies for their predictions. The first team had seemingly a dog expert and usually listened to that teammate's recommendation for predictions. The second team adopted an ensemble approach in which everyone independently made their predictions before discussing (in an effort to combat group think). The third team did all of their deliberation in Chinese because that was their native language, and was therefore easier for them to communicate.

Figure 5. Example training slide for team 2.

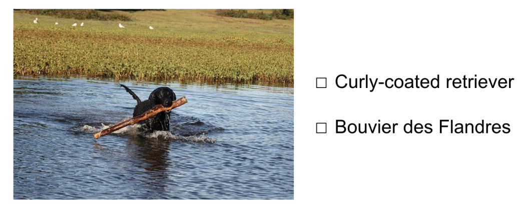


Figure 6. Prediction slide for the task from Figure 5

Figure 7. Perils of a correlation-based predictor being illustrated by humans. On showing images of curly-coated retrievers and Bouvier des Flanders, one of the teams' teachers described an incorrect *feature* to their team - of retrievers generally holding onto an object in their mouths. Coincidentally, the prediction image consisted of a retriever conveniently holding an object in its mouth, leading their team to guess that the good dog was a retriever. This demonstrates how learning a correlating feature which is not causal can have dire consequences in modern AI systems.

4 Results

Humans performed surprisingly well. All three teams got $\frac{7}{8}$(= 87.5%) predictions right. This is in the same range as human baseline results reported by Karpathy et al. for ImageNet [6]. Notes from all three teams are shown in Figures 8, 9, and 10.

Analyzing the notes revealed interesting information. One note from team 1 (not shown here) suggested that some teammates were more visually inclined than others, as they chose to draw the descriptions provided by their teacher rather than writing them out. This seems to have played no difference in the final accuracy though.

We can see in Figure 9 that the ensemble-based approach gives 6 times as many labels for the human baseline. Further, we can see that most team predictions were unanimous (6-0) or near unanimous (5-1), showing the predictions were not independent (and therefore all based in the same direction), thus defeating the noise cancelling benefits of ensembling.

5 Discussion

5.1 Dog descriptions

There were some very good human descriptions of the dogs, including:

1. "It's super cute and fluffy when it is small."
2. "The Corgi... looks like a dwarf. Like a dog meets a dwarf... in it's legs."
3. "They're very different. But each beautiful creatures in their own right."

5.2 Perils of correlation-based prediction systems

Occasionally, humans were arguably right for the wrong reasons. The teacher who was describing the classes in Figure 5 confidently stated that Curly-coated Retrievers "almost always are holding something in their mouths." Lo and behold, the prediction image (shown in Figure 6) showed a Curly-coated retriever holding something in its mouth. This suggests a potential downside to such feature-based, correlation-driven prediction systems. It is, of course, possible that this property is a true fact about the world. More research is required to establish its certainty.

5.3 Neuroscience of zero-shot transfer learning

This task opens up the possiblity of the existence of an information barrier between what is processed by the visual cortex, and how much of it can be retrieved by the language system in the human brain. In this task, the teacher observes images of the two breeds, processes it, and is required to then switch to using her/his language system to describe what they just perceived. Neuroscientifically, this requires the brain to distill a representation of the two breeds from the visual cortex, which is then supposedly accessed by the language system. In our tasks, we observe that each task from each team had less than a 100% accuracy, which suggests that the language system cannot potentially access all the information processed by the visual cortex. In a sense, there possibly exists an information barrier between the representation that is encoded by the visual cortex, and the information decoded from it by the language system. Such a lossy decoding is indicative of the brain using a low dimension feature space (as compared to the feature space enabled by retinal ganglion cells) to store such representations. As a caveat, the statistical validity of these results are limited ($n = 6$ subjects). This is a possible research question which can be tested through behavioral studies in the future.

6 Conclusion

All dogs are, and will remain good. Additionally, people are pretty good at identifying dogs without having seen them before. We speculate that assistive technology could help humans perform even better, thus allowing people to achieve super-human performance at identifying dogs. This could enable future citizens of the world to meet more dogs than ever imagined.

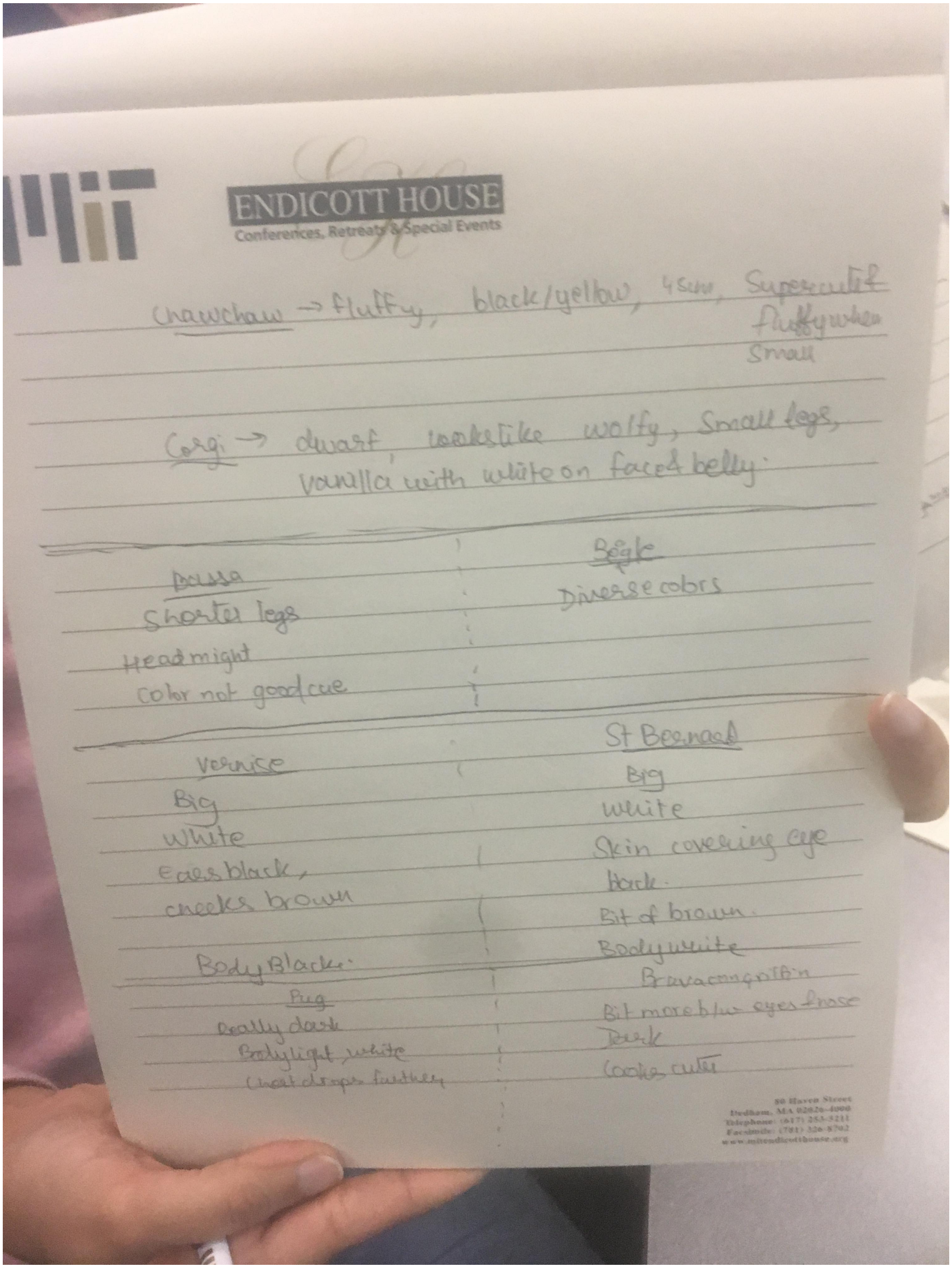
Acknowledgments

We thank the *ML Across MIT* organizing committee (especially Kyungmi Lee, Lei Xu, and Frederic Koehler) for putting this fun challenge together.

References

[1] W. Boag. 2018. Good Dogs of Cambridge. *SIG TBD* (2018). http://sigtbd.csail.mit.edu/pubs/2018/sigtbd18-paper-5.pdf
[2] @dog_rates. 2019. WeRateDogs. (2019). https://twitter.com/dog_rates
[3] Everyone. 2018. Doug the Pug. *Wikipedia* (2018). https://en.wikipedia.org/wiki/Doug_the_Pug
[4] M. Flynn. 2018. Monkey loses selfie copyright case. Maybe monkey should sue PETA, appeals court suggests. *Washington Post* (2018). https://www.washingtonpost.com/news/morning-mix/wp/2018/04/24/monkey-loses-selfie-copyright-case-maybe-monkey-should-sue-peta-appeals-court-suggests/?utm_term=.7c8ee5d997ef
[5] Kaggle. 2018. Dog Breed Identification: Determine the breed of a dog in an image. (2018). https://www.kaggle.com/c/dog-breed-identification
[6] Andrej Karpathy. 2014. What I learned from competing against a ConvNet on ImageNet. (2014). http://karpathy.github.io/2014/09/02/what-i-learned-from-competing-against-a-convnet-on-imagenet/
[7] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. 2011. Novel Dataset for Fine-Grained Image Categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*. Colorado Springs, CO.
[8] Play Until Dark LLC. 2012. Meet My Dog App. (2012). http://meetmydogapp.com/
[9] J.A. Lou. 2016. Dogs And Machine Learning Come Together In An App. *The Bark* (2016). https://thebark.com/content/dogs-and-machine-learning-come-together-app
[10] A. Ma. 2015. Tinder for Dogs. *Huffington Post* (2015). https://www.huffpost.com/entry/tindog-tinder-for-dogs_n_55aea38ee4b08f57d5d2c264
[11] M. Túlio Ribeiro, S. Singh, and C. Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR* abs/1602.04938 (2016). http://arxiv.org/abs/1602.04938
[12] Yichong Xu. 2019. Building Machine Learning Models via Comparisons. (2019). https://blog.ml.cmu.edu/2019/03/29/building-machine-learning-models-via-comparisons/

A Appendix


ENDICOTT HOUSE
Conferences, Retreats & Special Events

Chawchaw → fluffy, black/yellow, 4sun, Supercute
fluffy when
small

Corgi → dwarf, looks like wolfy, Small legs,
vanilla with white on face & belly.

Bassa
Shorter legs
Head might
Color not good cue

Beagle
Diverse colors

Vernise
Big
White
Ears black,
cheeks brown
Body Black.

St Bernard
Big
white
Skin covering eye
black.
Bit of brown.
Body white

Pug
Really dark
Bodylight, white
Cheat drops further

Bravacongriffin
Bit more blue eyes & nose
Dark
Looks cute

Figure 8. Notes taken by a member of team 1.

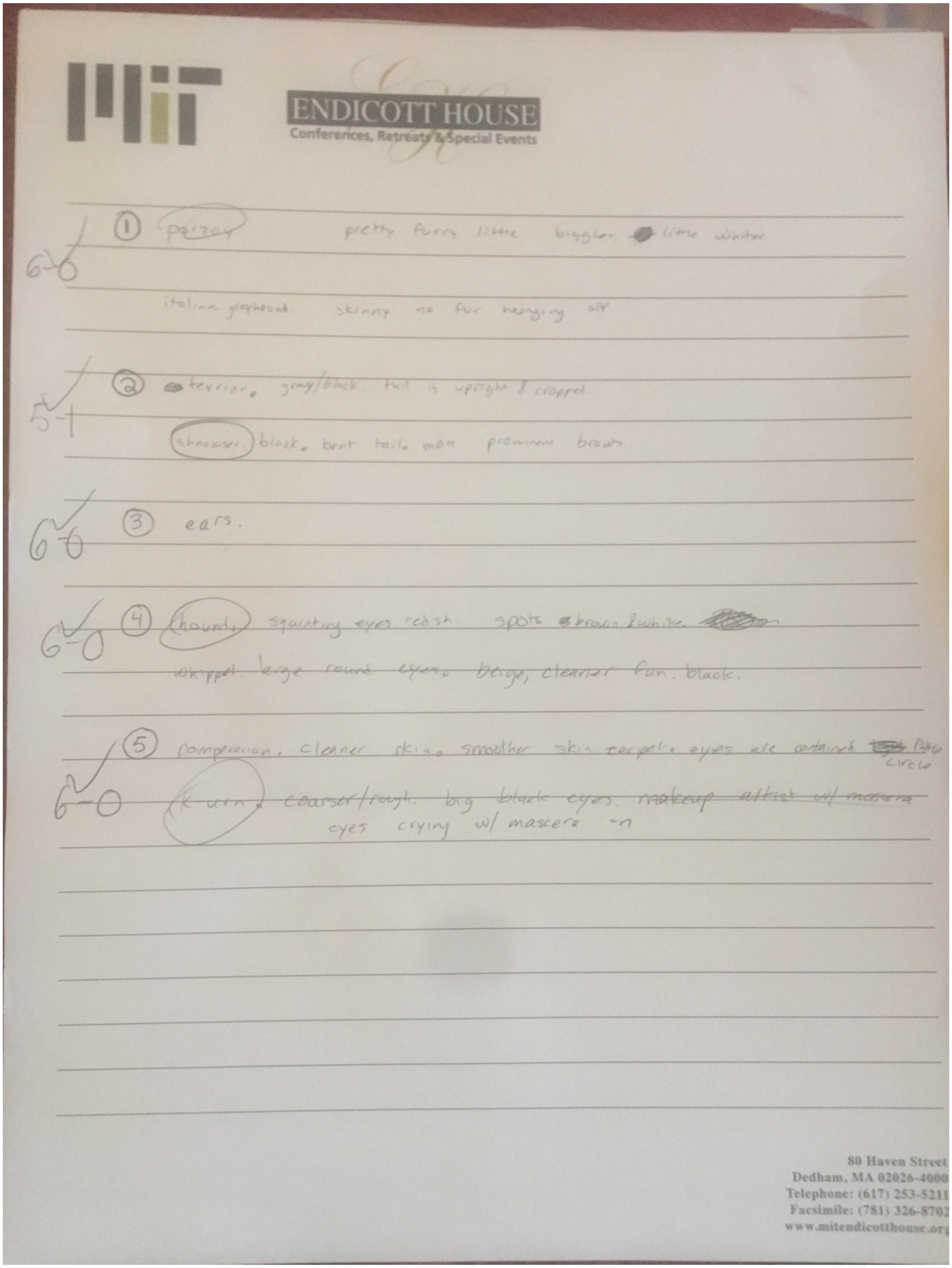
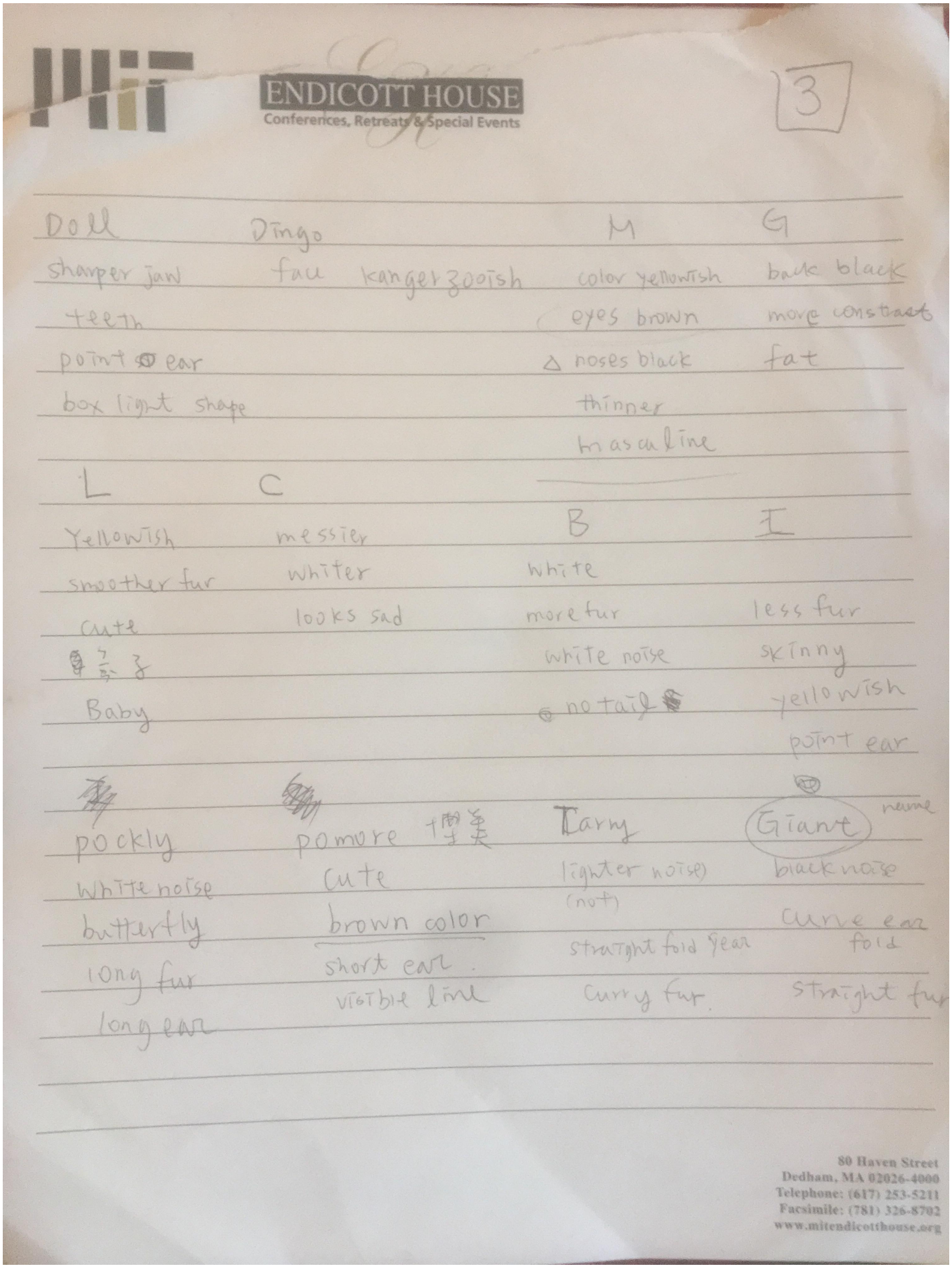


Figure 9. Notes taken by a member of team 2.



ENDICOTT HOUSE
Conferences, Retreats & Special Events

3

Doll
sharper jaw
teeth
point ear
box light shape

Dingo
fau kangerzooish

M
color yellowish
eyes brown
noses black
thinner
masculine

G
back black
more constrast
fat

L
Yellowish
smoother fur
cute
Baby

C
messier
whiter
looks sad

B
white
more fur
white noise
no tail

I
less fur
skinny
yellowish
point ear

pockly
White noise
butterfly
long fur
long ear

pomore
cute
brown color
short ear.
visible line

Larry
lighter noise)
(not)
straight fold year
curry fur.

Giant name
black noise
curve ear
fold
straight fur

80 Haven Street
Dedham, MA 02026-4000
Telephone: (617) 253-5211
Facsimile: (781) 326-8702
www.mitendicotthouse.org

Figure 10. Notes taken by a member of team 3.