Skip-Frame Embeddings for Feature Adaptation and Visualization

Zain Shah [1]

Abstract

We present an unsupervised method for visualizing the generalization and adaptation capabilities of pre-trained features on video. Like the skip-grams method for unsupervised learning of word vector representations, we exploit temporal continuity in the target media, namely that neighboring frames are qualitatively similar. By enforcing this continuity in the adapted feature space we can adapt features to a new target task, like house price prediction, without supervision. The domain-specific embeddings can be easily visualized for qualitative introspection and evaluation.

1. Introduction

1.1. Motivation

While deep learning techniques have shown impressive machine learning results in many problem domains, the data volume required for accurate generalization relative to more classical methods has made adaptation to less data-rich, industrial, and domain specific tasks difficult.

Transfer learning, a family of techniques where the knowledge gained from one task is adapted for another, presents a promising solution to this issue. With the proliferation of pre-trained models available online it is now possible to take advantage of existing work and regularities between tasks to achieve good performance on a different task than a model was originally trained on.

Specifically, one oft-used workflow is that of transfer learning for image classification in a specific domain. Instead of training a model from scratch one would fine-tune, or reuse the learned features of, a model trained to classify images on another task. Determining how much fine-tuning or further training will be necessary typically requires experimental verification, however. This impedes evaluation of multiple options, because gathering the labels required and even fine-tuning a pre-existing model on new data takes considerable effort.

[1]Opendoor Labs Inc., San Francisco, USA. Correspondence to: Zain Shah <zain.shah@opendoor.com>.

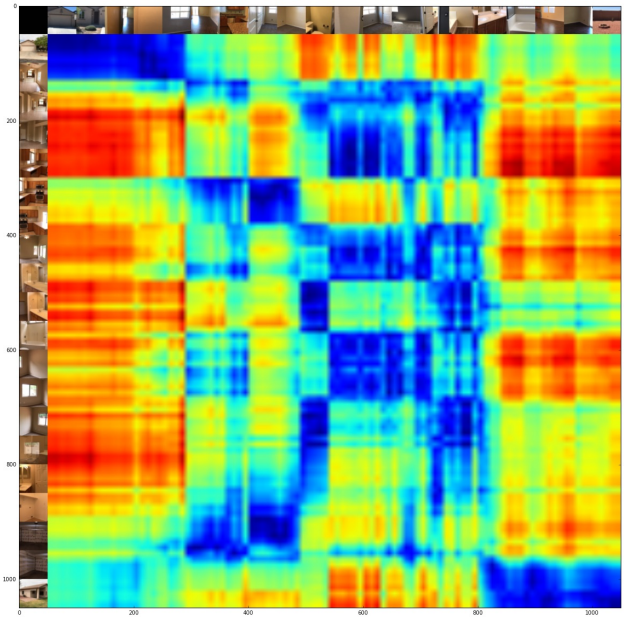

Figure 1. Two home video walkthroughs compared with Skip-Frame embeddings visualized in a pairwise distance matrix. Blue regions are low distance, and here correspond to learned semantic alignment of the two videos.

1.2. Our approach

In this paper we present an unsupervised technique that proves useful for transfer learning via pre-trained feature adaptation, and is especially useful for evaluating the generality of existing features to a new target task. First, we recognize that time series signals, like audio and video, present temporal regularities that are useful for unsupervised feature learning.

In a video, for example, temporally adjacent frames are likely to be more visually similar than frames from a different part of the video. In audio, this same continuity is likely. By training a shallow embedding model with a max-margin ranking loss to encourage samples of media that are temporally adjacent to be near one another in the embedding space, we can enforce this regularity in the adapted feature space. While there are some discontinuities in the media, our method effectively includes data augmentation because

it requires many temporally distant negative samples, and thereby appears robust to such discontinuities.

The outline of the paper is as follows. In Section 2 we outline the Skip-Frame method, its implementation, and its relation to Skip-grams. In Section 3 we present an application of the method to a domain specific task, namely nearest-neighbor regression of house prices from video walkthroughs. In Section 4 we visualize the learned feature representations and learn how to evaluate their generalization capabilities and robustness to discontinuities. Subsequently, Section 5 presents potential future directions of research and our conclusions.

2. Approach

2.1. Related Work

The technique we present here for learning Skip Frame embeddings is most closely related to Mikolov, et. al.'s Skip-grams technique (2013a) used for unsupervised learning of generic word vector representations from a large corpus. While the original skip-grams objective is to find word representations that are useful for predicting the surrounding words in a sentence or a document, we additionally include negative samples in our approach, analogous to the negative sampling developed further in (Mikolov et al., 2013b) and previously explored as a form of Noise Contrastive Estimation (Mnih & Whye Teh, 2012).

Specifically, we follow the work of (Collobert & Weston, 2008; Kiros et al., 2015) in using a hinge loss objective to encourage temporally adjacent frames to be ranked closer to one another than discontinuous negative frame pairs (as such, our technique could be considered a form of single-view time contrastive embedding, as explored in (Sermanet et al., 2017)).

2.2. Skip-Frames objective

Suppose we have a set of pre-trained features on the target media, but not the target task or domain, $X \in \mathbb{R}^{N \times D}$ for N time-series with dimensionality D. Let X_i be the set of features for time-series i, where $x_i^t \in \mathbb{R}^D$ denotes the D-dimensional feature vector at time t and series i.

Our training data consists of a batch of triples, with randomly sampled anchor frames a_i^t, positive frames p_i^a within a neighboring time margin T, and negative frames n_i^a outside of that time margin. We want to learn an embedding over X with embedding matrix $\mathbf{U}$, s.t.

$$||\mathbf{U}p_i^a - \mathbf{U}a_i^t|| < ||\mathbf{U}n_i^a - \mathbf{U}a_i^t|| \tag{1}$$

Thereby, we seek to optimize the following pairwise-ranking loss:

Figure 2. Example triples the model sees during training. A subsampled, tiled, sequence of frames in the video is centered for convenience, with two example triples to its left and right.

$$\min_{\boldsymbol{U}} \sum_{\mathbf{i}} \max\{0, \lambda - ||\mathbf{U}p_i^a - \mathbf{U}a_i^t|| + ||\mathbf{U}n_i^a - \mathbf{U}a_i^t||\}$$

In the above formulation of the loss we do not specify a distance metric - any distance function (e.g. euclidean, cosine, etc.) can be used, though we recommend using cosine distance if the feature magnitudes in X are not $X \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$.

The free parameters to be learnt are in the embedding matrix $\mathbf{U}$ with dimensionality $D \times d$ where D is the dimensionality of the source features and d is the desired target dimensionality. The only additional parameters are the margin λ for which we used $\lambda = 0.05$, and for numerical stability purposes we replace 0 with 1×10^{-6}.

3. Experiments

We report results and analysis from adapting pre-softmax feature vectors from a variety of models trained on ImageNet to the task of predicting house price valuation from video walkthroughs. We report best results using our technique to adapt the features of the Inception ResNet50 architecture, with PReLu activations applied to our embedding matrix $\mathbf{U}$ and a dropout probability of 0.2.

3.1. House Price Prediction Task

We consider the task of accurately predicting the prices of a set of homes given only 2 features - the home's address, and a video walk-through of the home. This task is of interest because typical home valuation models require a bevy of hand-engineered quantitative features which neither fully characterize the home nor present a positive user experience for anyone interacting with the model, either by providing the data or trying to investigate its results. On the other hand, a single video walk-through of a home conveys most of the qualitative aspects of a home, except for its location, and any predictions made from this data are easily interpretable by a layperson.

Typically to solve such a task we would need to already have a large, labeled dataset of $\{video, location, price\}$ triples. In many industry cases, including this one, evaluating the viability of learning such a task must preclude the cost of gathering the requisite data. Thus, we utilize Skip-Frames to take advantage of the relatively small dataset we do have, $\mathcal{O}(100)$, to learn a nearest-neighbors regression in an unsupervised manner.

3.2. Dataset

Our proprietary dataset contains 165 {video, location, close price} triples of homes in Phoenix, Arizona. The videos are all 640x320 resolution, 30fps, RGB, and vary in duration from 1 to 2 minutes. The locations are specified by street address, which we geocode into latitude, longitude using the Google Maps API. We also augment the available data by taking crops of the frames bordering each corner at 224x224 resolution before computing their generic visual features. After data augmentation and sampling negative and positive frame pairs, our dataset yields $30\,fps * 120\,seconds * 165\,videos * 3\,samples = 1,782,000\,samples$ to train from.

To portray our task we make a mosaic of a random sample of the dataset available on YouTube.

3.3. Training

Training proceeds as described in 2.2. From the videos we sample temporally adjacent positive frames, and outside of that margin (5 seconds yielded best results in our experiments) we sample negative frames. A set of such triples are illustrated in Figure 2. We utilize the objective on frame level pre-softmax features from both the Oxford VGGNet 16 layer CNN and the Inception ResNet 50 layer CNN. Our final model consists of 2 PReLU layers with 0.2 dropout probability applied to the frame level features during training, followed by a single affine layer projecting the hidden layer features into our final embedding space.

3.4. Visualization and Pre-Evaluation

Prior to evaluation on the final task, we want to understand exactly what we learn with this objective, and the extent to which these embeddings generalize between examples. To do so we take advantage of the fact that the distances enforced by the objective function can be visualized in a pairwise distance matrix. For 2 videos (either of the same home or different, with lengths N and M) we form a distance matrix $D \in \mathcal{R}^{\mathcal{N}\times\mathcal{M}}$ where each element d_m^n is the distance between feature vector x_n and x_m. This matrix can be visualized as a heatmap, where the time series M and N are displayed along the edges of the heatmap, as in Figures 1, 4 & 5.

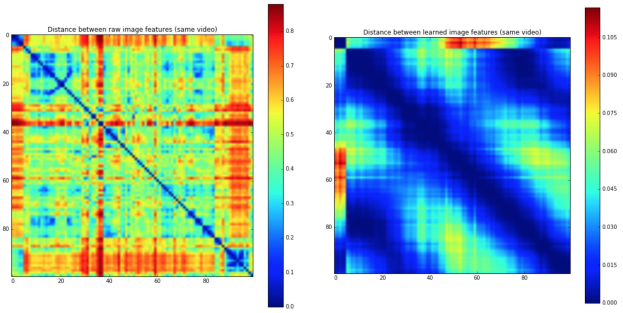

Figure 3. On the left we show the pairwise distance matrix between raw image features from Inception ResNet 50, vs. our learned embeddings on the right. The colormap goes from blue to red, low to high cosine distance. The widened diagonal blue line shows our embeddings learn to generalize beyond exact visual matches.

Such pairwise distance heatmaps between frame level embeddings can be visually inspected quite easily. If N and M are the same video, then effective embeddings have a region of low distances that generalizes beyond the exact diagonal - that is, we expect frames which are temporally neighboring but not at exactly the same time to still be near one another, as specified by our objective. We can see this difference between the raw features, embeddings learned from VGG16, and embeddings learned from ResNet50 quantitatively in Table 1 and visually in Figures 4 & 5.

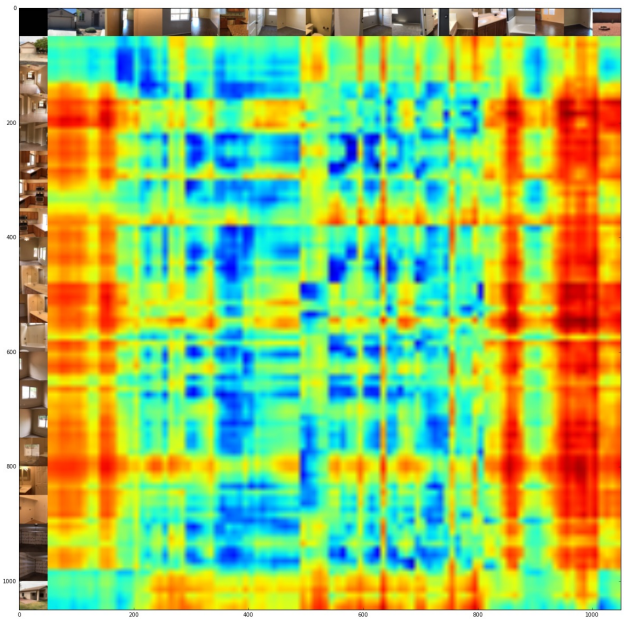

Figure 4. Two home video walkthroughs compared with Skip-Frame embeddings learned on VGG features visualized in a pairwise distance matrix. Here we can see no strong semantic alignment - evidence that the VGG features are less effective than the Inception features.

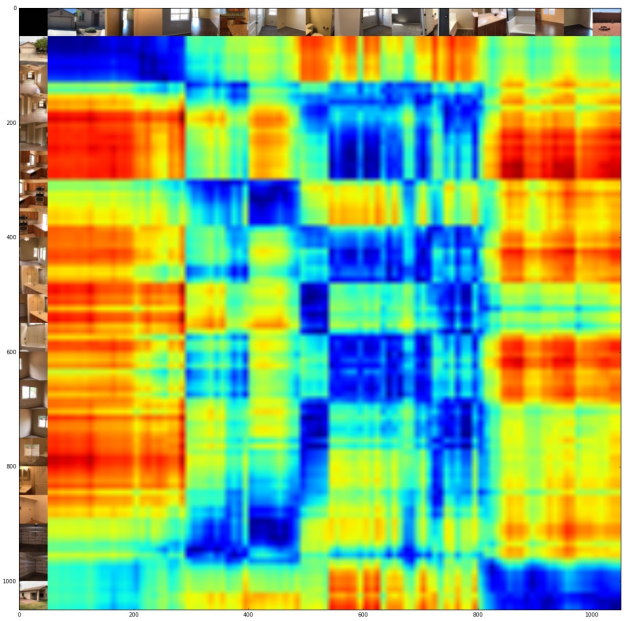

Figure 5. Two home video walkthroughs compared with Skip-Frame embeddings learned from Inception ResNet 50 features, visualized in a pairwise distance matrix. Blue regions are low distance and here correspond to learned strong semantic alignment of the two walkthroughs.

Additionally, we evaluate how well our objective generalizes to comparisons between frames of **different** homes, by using different homes for M and N, then visualizing the mean distance matrix as in Figure 5. There we see how the same wide diagonal we expect between frames of the same video also exists between videos (the subject matter is filmed in roughly the same order every time, so this is to be expected).

For individual examples, we can investigate which characteristics of the video the embeddings consider similar. In Figure 5 for example, it is apparent from the image that footage of the interior and exterior is easy to distinguish, while kitchens and bathrooms are more difficult to distinguish.

3.5. Evaluation

Because neither the similarity between homes nor the final price objective that we are evaluating our performance on present themselves in our objective function, we evaluate our performance on the entire dataset by devising a slightly nontraditional physical-distance weighted k-nearest-neighbors regression model from the pairwise distance matrices above. For a given home, we compute the mean of the pairwise distance matrix between that home and every other home, in addition to the physical distance between the homes, and apply a sigmoid and then softmax to both of those features. Finally a softmax is applied across the normalized, weighted distances between homes, and that distance vector is used to compute a weighted sum of the prices of each home (except the target home) to arrive at a prediction. The results of this prediction, compared to using the raw ResNet50 features and a naive baseline (predicting the mean price) are presented in Table 1. The Skip-Frames objective allows us to predict prices for the homes in our dataset with a Median Absolute Error of about 13%, compared to 15% and 16% for our baselines, respectively.

Table 1. Price regression Mean Absolute Error (MAE) for various models, vs lowest validation loss on our Skip-Frames objective. Our embeddings are labeled **SF**. Baseline is simply predicting the mean price.

MODEL	MAE	LOSS
VGG-16 (SF)	15%	0.0305
INCEPTION (SF)	13%	0.0235
INCEPTION (RAW)	17%	-
BASELINE	18%	-

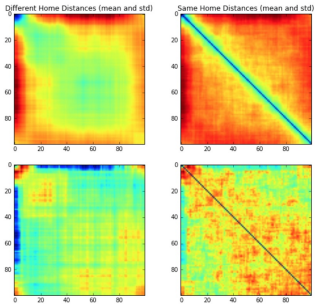

Figure 6. We can compare how well our embeddings generalize across homes by visualizing the pairwise distance matrices for different homes. In the left column we show the mean (top) and variance (bottom) for pairwise distance matrices of different homes. On the right we show the mean and variance of distance matrices for the same homes.

4. Conclusion

This paper presents a technique for visualizing and adapting features from a pre-trained model for a new out of domain task. Consider our initial motivation - in most prediction and classification use cases in industry viability of the prediction task precludes the costly gathering of data. By exploiting the regularities in time-series data, we present a method for validating the gains to be made in learning features specific to a domain/task in a fully unsupervised manner, and visualizing those learned features.

In future work, we plan to explore the applicability of this technique to other media, such as audio, and investigate the potential for learning cross-modal regularity between audio and imagery of the same video. In such cases, the directly interpretable visualizations lend themselves well to interactive data exploration.

References

Collobert, Ronan and Weston, Jason. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 160–167, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-205-4. doi: 10.1145/1390156.1390177. URL http://doi.acm.org/10.1145/1390156.1390177.

Kiros, Ryan, Zhu, Yukun, Salakhutdinov, Ruslan, Zemel, Richard S., Torralba, Antonio, Urtasun, Raquel, and Fidler, Sanja. Skip-thought vectors. *CoRR*, abs/1506.06726, 2015. URL http://arxiv.org/abs/1506.06726.

Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013a. URL http://arxiv.org/abs/1301.3781.

Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S, and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013b. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phr pdf.

Mnih, A. and Whye Teh, Y. A Fast and Simple Algorithm for Training Neural Probabilistic Language Models. *ArXiv e-prints*, June 2012.

Sermanet, Pierre, Lynch, Corey, Hsu, Jasmine, and Levine, Sergey. Time-contrastive networks: Self-supervised learning from multi-view observation. *CoRR*, abs/1704.06888, 2017. URL http://arxiv.org/abs/1704.06888.