Rapid Image Annotation Through Zero-Shot Learning

Abstract

Recent experiments on word analogies demonstrate that contemporary word vectors effectively encapsulate subtle linguistic patterns through linear vector displacements. However, the extent to which these straightforward vector displacements can represent visual patterns across words remains uncertain. This research investigates a particular image-word relevance relationship. The findings indicate that, for a given image, word vectors of pertinent tags are positioned higher than those of unrelated tags along a primary axis within the word vector space. Drawing inspiration from this insight, we suggest addressing image tagging by determining the main axis for an image. Specifically, we utilize linear mappings and intricate deep neural networks to deduce the primary axis from an input image. The resultant tagging model exhibits remarkable adaptability. It operates swiftly on test images, with a processing time that remains constant regardless of the training set's size. Furthermore, it showcases exceptional performance not only in conventional tagging tasks using the NUS-WIDE dataset but also in comparison to competitive baselines when assigning tags to images that haven't been seen during training.

1 Introduction

Recent advancements in representing words in vector spaces have proven advantageous for both Natural Language Processing and various computer vision applications, including zero-shot learning and image caption generation. The rationale behind using word vectors in NLP is rooted in the observation that detailed linguistic patterns among words are represented by linear offsets of word vectors. This pivotal insight emerged from well-known word analogy studies. For example, syntactic relationships like "dance" to "dancing" parallel "fly" to "flying," and semantic connections like "king" to "man" mirror "queen" to "woman." Nevertheless, it is yet to be determined whether the visual patterns across words, implicitly employed in the aforementioned computer vision tasks, can similarly be represented by these basic vector offsets.

This paper focuses on the task of image tagging, where an image necessitates the division of a word lexicon into two distinct groups based on image-word relevance. For example, an image of a zoo might have relevant tags like "people," "animal," and "zoo," while irrelevant tags might include "sailor," "book," and "landscape." This lexical division fundamentally differs from the nuanced syntactic or semantic relationships examined in word analogy tests. Instead, it concerns the connection between two sets of words as prompted by a visual image. This type of word relationship is semantic and descriptive, emphasizing visual association, albeit at a broader level. Given this context, it is worth investigating whether word vectors maintain the property where simple linear vector offsets can depict visual or image-based associative relationships between words. In the zoo example, while it's easy for humans to recognize that words like "people," "animal," and "zoo" are more related to the zoo than words like "sailor," "book," and "landscape," the question is whether such a zoo-association relationship can be represented by the nine pairwise vector offsets: "people" minus "sailor," "people" minus "book," and so on, up to "zoo" minus "landscape," between the vectors of relevant and irrelevant tags.

A primary contribution of this research is an empirical investigation of these questions. Each image establishes a visual association rule over words, represented as a pair (Y, Y). Leveraging the extensive

.

image collections in benchmark datasets designed for image tagging, we can explore numerous distinct visual association rules in words and the corresponding vector offsets in the word vector space. Our findings uncover a significant correlation: the offsets between the vectors of relevant tags (Y) and those of irrelevant tags (Y) predominantly align in a consistent direction, which we term the "principal direction". In other words, within the word vector space, there exists at least one vector (direction), denoted as w, such that its inner products with the vector offsets between Y and Y are greater than 0. This can be expressed as:

(w,p Ž014 n) > 0 equivalently, (w,p) > (w,n)

This implies that the vector w ranks all relevant words Y ahead of irrelevant ones Y.

The visual association patterns among words manifest as the linear rank-abilities of their corresponding word vectors. This observation corroborates findings from word analogy studies, suggesting that multiple relationships for a single word are embedded within a high-dimensional space. Furthermore, these relationships can be articulated using basic linear vector arithmetic.

Building on this discovery, we propose a solution to the image tagging challenge by identifying the primary axis along which relevant tags are ranked higher than irrelevant ones within the word vector space. We employ both linear mappings and deep neural networks to infer this primary axis from each input image. This unique perspective on image tagging yields a highly adaptable tagging model. The model processes test images rapidly, maintaining a constant processing time irrespective of the training dataset's size. It not only delivers outstanding results in traditional tagging tasks but also excels at assigning new tags from a broad vocabulary that were not encountered during training. Our method does not rely on prior knowledge of these new tags, as long as they exist within the same vector space as the tags used during training. Consequently, we designate our technique as "fast zero-shot image tagging" (Fast0Tag), acknowledging its strengths in both speed and its zero-shot learning capabilities.

In stark contrast to our approach, prior methods for image tagging are limited to assigning only those tags to test images that were seen during training, with a notable exception. These methods are constrained by the fixed and often limited number of tags present in the training data, which poses practical challenges. For example, Flickr hosts approximately 53 million tags, and this number is rapidly increasing. The work of Fu et al. represents a pioneering effort to extend an image tagging model to previously unseen tags. However, when compared to our proposed method, it depends on two extra assumptions. Firstly, it assumes that unseen tags are known beforehand to enable model adjustment toward these tags. Secondly, it assumes that test images are known in advance for model regularization. Moreover, this method is restricted to a very limited number, U, of unseen tags, as it needs to account for all $2^U possibletagcombinations$.

To recap, our primary contribution lies in analyzing visual association patterns in words as they relate to images and how these patterns are reflected in word vector offsets. We posit and confirm through experiments that a main direction exists in the word vector space for each visual association rule (Y, Y), where vectors of relevant words are ranked higher than others. Building on this, our second contribution is an innovative image tagging model, Fast0Tag, which is both swift and capable of handling an open vocabulary of unseen tags. Lastly, we explore three distinct image tagging scenarios: traditional tagging, which assigns seen tags to images; zero-shot tagging, which annotates images with numerous unseen tags; and seen/unseen tagging, which uses both seen and unseen tags. Existing research either addresses traditional tagging or zero-shot tagging with a limited number of unseen tags. Our Fast0Tag method surpasses competitive baselines across all three scenarios.

2 Related Work

Image Tagging. The objective of image tagging is to allocate pertinent tags to an image or to generate a ranked list of tags. Within the academic community, this challenge has predominantly been tackled from the standpoint of tag ranking. Generative approaches, which incorporate topic models and mixture models, inherently rank candidate tags based on their conditional probabilities relative to the test image. Conversely, non-parametric, nearest-neighbor-based techniques frequently rank tags for a test image by aggregating votes from a selection of training images. Although nearest-neighbor methods generally exhibit superior performance compared to those reliant on generative models, they are plagued by substantial computational demands during both training and testing phases.

The recently introduced FastTag algorithm offers a significant speed advantage while maintaining performance levels on par with nearest-neighbor methods. Our Fast0Tag method mirrors the reduced complexity of FastTag. Embedding techniques, on the other hand, determine tag ranking scores via a cross-modal mapping between images and tags. This concept has been further developed using deep neural networks. Notably, aside from certain exceptions, the majority of these methods do not train their models with an explicit ranking objective, despite ultimately ranking candidate tags for test images. This discrepancy between the trained models and their practical application contravenes the principle of Occam's razor. We incorporate a ranking loss in our approach, similar to these exceptions.

Unlike our Fast0Tag, which is capable of ranking both known and an unlimited number of previously unseen tags for test images, the methods mentioned earlier are restricted to assigning tags to images from a predetermined vocabulary encountered during training. An exception to this is the work by Fu et al., where they address a predefined number, U, of unseen tags by developing a multi-label model that considers all possible $2^U combinations of these etags. However, this approach is constrained by the small number U of unseen tags it can handle.$

Word Embedding. Diverging from the conventional one-hot vector representation of words, word embedding maps each word to a continuous-valued vector, primarily learning from the statistical patterns of word co-occurrences. While earlier studies on word embedding exist, our research emphasizes the latest GloVe and word2vec vectors. As demonstrated in the well-known word analogy experiments, both types of word vectors effectively capture detailed semantic and syntactic patterns through vector offsets. In this study, we further reveal that basic linear offsets can also represent the broader visual association patterns among words.

Zero-Shot Learning. The term "zero-shot learning" is frequently used interchangeably with "zero-shot classification," although the latter is actually a subset of the former. In contrast to weakly-supervised learning, which acquires new concepts by extracting information from noisy samples, zero-shot classification aims to classify objects from unseen classes by learning classifiers from seen classes. Attributes and word vectors are two primary semantic sources that enable zero-shot classification.

Our Fast0Tag, together with Fu et al., expands the domain of zero-shot learning to include zero-shot multi-label classification. Fu et al. approach this by converting the problem into zero-shot classification, where each combination of multiple labels is treated as a separate class. We, on the other hand, model the labels directly, allowing us to assign or rank a large number of unseen tags for an image.

3 The Linear Rank-Ability of Word Vectors

Our Fast0Tag method is enhanced by the discovery that the visual relationship between words, specifically how a lexicon is divided based on relevance to an image, manifests in the word vector space as a main direction. Along this direction, words or tags that are relevant to the image are ranked higher than those that are not. This section elaborates on this discovery.

3.1 The Regulation Over Words Due to Image Tagging

Let's denote S as the set of seen tags available for training image tagging models, and U as the set of tags unseen during the training phase. The training data is structured as (xm, Ym); m = 1, 2, ..., M, where xm represents the feature vector of image m in RD, and Ym is a subset of S, containing the seen tags relevant to that image. For simplicity, we also use Ym to represent the collection of corresponding word or tag vectors.

Traditional image tagging seeks to assign seen tags from S to test images. Zero-shot tagging, as defined by Fu et al., aims to annotate test images using a predetermined set of unseen tags, U. Beyond these two scenarios, this paper introduces seen/unseen image tagging, which identifies both relevant seen tags from S and relevant unseen tags from U for test images. Furthermore, the set of unseen tags, U, can be open and continuously expanding.

We define Ym as the complement of Ym in S, representing irrelevant seen tags. An image m establishes a visual association rule among words, essentially partitioning seen tags into two distinct sets: Ym and Ym. Recognizing that various detailed syntactic and semantic patterns among words

can be depicted through linear word vector offsets, we proceed to investigate the characteristics these vector offsets might exhibit for this novel visual association rule.

3.2 Principal Direction and Cluster Structure

Figure 2 offers a visual representation of vector offsets (p - n), where p belongs to Ym and n belongs to Ym, using both t-SNE and PCA for two different visual association rules over words. One rule is defined by an image associated with 5 relevant tags, and the other by an image with 15 relevant tags. From these vector offsets, we identify two key structures:

Principal Direction: For a given visual association rule (Ym, Ym) in words for image m, the vector offsets predominantly point in a similar direction, which we refer to as the principal direction. This suggests that along this principal direction, relevant tags Ym are ranked higher than irrelevant ones Ym.

Cluster Structure: Within each visual association rule over words, there are discernible cluster structures in the vector offsets. Moreover, all offsets that point to the same relevant tag in Ym are grouped within the same cluster. In Figure 2, we distinguish offsets pointing to different relevant tags by using different colors.

The question remains whether these two observations can be generalized. Specifically, do they remain valid in the high-dimensional word vector space for a broader range of visual association rules defined by other images? To address this, we designed an experiment to confirm the existence of principal directions in word vector spaces, or equivalently, the linear rank-ability of word vectors. We defer the investigation of the cluster structure to future research.

3.3 Testing the Linear Rank-Ability Hypothesis

The experiments in this section are performed using the validation set of the NUS-WIDE dataset, which includes 26,844 images, 925 seen tags (S), and 81 unseen tags (U). The number of relevant seen/unseen tags associated with an image varies from 1 to 20/117, with an average of 1.7/4.9. Further details can be found in Section 5.

Our goal is to explore whether a primary direction exists for any visual association rule (Ym, Ym) created by image m, along which relevant tags Ym rank higher than irrelevant tags Ym. This can be confirmed if we find a vector w in the word vector space that fulfills the ranking conditions (w, p) > (w, n) for all p in Ym and n in Ym.

To achieve this, we train a linear ranking SVM for each visual association rule using all corresponding pairs (p, n). We then rank word vectors using the SVM and assess the number of violated constraints. Specifically, we use MiAP, with higher values being preferable, to compare the SVM's ranking list against the ranking constraints. This process is repeated for all validation images, resulting in 21,863 unique visual association rules.

Ranking SVM Implementation. We utilize the primal formulation of ranking SVM for our experiments, which is defined as:

min 1/2 ||w||$^2 + max(0, 1 - (w, yi) + (w, yj))foryiYm, yjYm$

Here, is a hyperparameter that balances the objective and regularization.

Results. The average MiAP outcomes across all distinct regulations are presented in Figure 3(left). We evaluate 300D GloVe vectors and word2vec vectors of dimensions 100, 300, 500, and 1000. The horizontal axis represents various regularizations used for training the ranking SVMs, with higher values indicating stronger regularization. In the 300D GloVe space and word2vec spaces of 300, 500, and 1000 dimensions, more than two ranking SVMs, with low values, produce nearly ideal ranking results (MiAP 1). This demonstrates that seen tags S are linearly rankable under almost every visual association rule, satisfying all ranking constraints set by relevant Ym and irrelevant Ym tags for image m.

However, caution is advised before extending conclusions beyond the experimental vocabulary S of seen tags. While an image m imposes a visual association rule over all words, this rule leads to different partitions of distinct experimental vocabularies (e.g., seen tags S and unseen tags U).

Therefore, we anticipate that the principal direction for seen tags should also apply to unseen tags under the same rule, if the questions at the end of Section 3.2 are answered affirmatively.

Generalization to Unseen Tags. We investigate whether the same principal direction applies to both seen and unseen tags under each visual association rule induced by an image. This is partially validated by applying the previously trained ranking SVMs to unseen tag vectors, as the "true" principal directions are unknown. We use the 81 unseen tags U as "test data" for the trained ranking SVMs, each resulting from an image-induced visual association. NUS-WIDE provides annotations for these 81 tags. The results, shown in Figure 3(right), significantly outperform the basic baseline of random tag ranking, indicating that the directions produced by SVMs are generalizable to the new vocabulary U of words.

Observation. We conclude that word vectors are an effective medium for transferring knowledge—specifically, rank-ability along the principal direction—from seen to unseen tags. We have empirically confirmed that the visual association rule (Ym, Ym) in words due to an image m can be represented by the linear rank-ability of corresponding word vectors along a principal direction. Our experiments involve a total of |S| + |U| = 1,006 words. Future work should include larger-scale and theoretical studies.

4 Approximating the Linear Ranking Functions

This section introduces our Fast0Tag approach for image tagging. Initially, we explain how to address image tagging by approximating the principal directions, based on their existence and generalization, as confirmed in the previous section. Subsequently, we describe the detailed approximation methods used.

4.1 Image Tagging by Ranking

Based on the findings from Section 3, which indicate the existence of a principal direction, wm, in the word vector space for each visual association rule (Ym, Ym) generated by an image m, we propose a direct solution for image tagging. The core idea is to approximate this principal direction by learning a mapping function, f(Ŏ0b7), that connects the visual space to the word vector space, such that:

f(xm) wm

Here, xm is the visual feature representation of image m. Consequently, given a test image x, we can promptly suggest a list of tags by ranking the word vectors of the tags along the direction f(x), specifically by the ranking scores:

t S U, (f(x), t)

This applies whether the tags are from the seen set S or the unseen set U.

We investigate both linear and nonlinear neural networks to implement the approximation function f(x) w.

4.2 Approximation by Linear Regression

In this approach, we assume a linear function from the input image representation x to the output principal direction w, defined as:

f(x) := Ax

Here, A can be determined in a closed form through linear regression. Thus, from the training data, we have:

$w_m = Ax_m + _m, form = 1, 2, ..., M$

where $w_m istheprincipaldirectionforalloffsetvectorsoftheseentags, correspondingtothevisualassociationrule(Y_m,$
$formsolutionforA.$

However, a challenge arises as we do not know the exact principal directions $w_m.Thetrainingdataonlyprovideimagesx_mandrelevanttagsY_m.Weoptforastraightforwardalternative, usingthedi$
$Ax.Thefirststagetrainsa rankingSVMoverthewordvectorsofseentagsforeachvisualassociation(Y_m, Y_m).Theseco$

Discussion. The use of linear transformation between visual and word vector spaces has been previously explored, for instance, in zero-shot classification and image annotation/classification. This work distinguishes itself by the clear interpretation of the mapped image f(x) = Ax as the principal direction for tag assignment, which has been empirically validated. We further extend this to a nonlinear transformation using a neural network.

4.3 Approximation by Neural Networks

We also explore a nonlinear mapping f(x;) using a multi-layer neural network, where represents the network parameters. The network architecture, illustrated in Figure 4, includes two RELU layers followed by a linear layer that outputs the approximated principal direction, w, for an input image x. We anticipate that the nonlinear mapping function f(x;) will provide greater modeling flexibility compared to the linear approach.

Training the neural network by regressing to the M directions obtained from ranking SVMs is not ideal, as confirmed by both intuition and experiments. The number of training instances, M, is small relative to the network's parameter count, increasing the risk of overfitting. Moreover, the directions from ranking SVMs are not the true principal directions, making it unnecessary to rely on them.

Instead, we integrate the two stages from Section 4.2. We aim for the neural network's output f(xm;) to represent the principal direction, where all relevant tag vectors p Ym rank higher than irrelevant ones n Ym for an image m. Let's define:

v(p, n;) = (f(xm;), n) - (f(xm;), p)

as the degree of violation of these ranking constraints.

We then minimize the following loss function to train the neural network:

* = argmin $w_m * l(x_m, Y_m;)l(x_m, Y_m;) = log(1 + expv(p, n;))forpY_m, nY_m$

where $w_m = 1/(|Y_m|*|Y_m|)normalizestheper-imageRankNetlossbythenumberofrankingconstraintsimposedbyima$
$batchgradientdescent.$

Practical Considerations. We use Theano for optimization, with a mini-batch size of 1,000 images. Each image, on average, imposes 4,600 pairwise ranking constraints, which are all used in the optimization. The normalization $w_m fortheper - imagerankinglosshelpsbalancetheinfluenceofimageswithmanypositivetags, addressingtheissueofunbalancednum$

Besides the RankNet loss, we tested other per-image loss options, including hinge loss, Crammer-Singer loss, and pairwise max-out ranking. Hinge loss performed the worst, likely because it's not designed for ranking. Crammer-Singer, pairwise max-out, and RankNet yielded comparable results, with RankNet slightly outperforming the others by about 2% in MiAP, possibly due to easier optimization control. Listwise ranking loss could also be considered.

5 Experiments on NUS-WIDE

This section details our experimental results, comparing our method against several strong baselines for traditional image tagging on the large-scale NUS-WIDE dataset. Additionally, we evaluate our method on zero-shot and seen/unseen image tagging scenarios, extending some existing zero-shot classification algorithms and exploring variations of our approach for comparison.

5.1 Dataset and Configuration

NUS-WIDE Dataset. We primarily utilize the NUS-WIDE dataset for our experiments. This dataset is a standard benchmark for image tagging, originally containing 269,648 images. We were able to retrieve 223,821 images, as some were either corrupted or removed from Flickr. Following the recommended protocol, we divide the dataset into a training set of 134,281 images and a test set of 89,603 images. We further allocate 20% of the training set as a validation set for tuning hyperparameters in both our method and the baselines, and for conducting the empirical analyses in Section 3.

Annotations of NUS-WIDE. NUS-WIDE provides three sets of tags for its images. The first set includes 81 "ground truth" tags, carefully selected to represent Flickr tags, encompassing both general terms (e.g., "animal") and specific ones (e.g., "dog," "flower"), and corresponding to frequent Flickr tags. These tags are annotated by students and are less noisy than those directly collected from the Web, serving as the ground truth for evaluating image tagging methods. The second and third sets contain 1,000 popular and nearly 5,000 raw Flickr tags, respectively.

Image Features and Word Vectors. We extract and normalize image feature representations using VGG-19. Both GloVe and Word2vec word vectors are used in our empirical analysis in Section 3, with 300D GloVe vectors used for the remaining experiments. Word vectors are also normalized.

Evaluation. We assess tagging results using two types of metrics: mean image average precision (MiAP), which considers the entire ranking list, and precision, recall, and F1-score for the top K tags in the list (K = 3 and K = 5). Both metrics are commonly used in image tagging research. For details on calculating MiAP and top-K precision and recall, we refer readers to Section 3.3 of Li et al. (2015) and Section 4.2 of Gong et al. (2013), respectively.

5.2 Conventional Image Tagging

In this section, we present experimental results for traditional image tagging, using the 81 "ground truth" annotated concepts in NUS-WIDE to benchmark various methods.

Baselines. We include TagProp as a primary competitive baseline, representing nearest-neighbor-based methods that generally outperform parametric methods built from generative models and have shown state-of-the-art results in experimental studies. We also compare against two recent parametric methods, WARP and FastTag, both based on deep architectures but using different models. For a fair comparison, we use the same VGG-19 features across all methods, with code for TagProp and FastTag provided by the authors and WARP implemented based on our neural network architecture. Additionally, we compare to WSABIE and CCA, which correlate images and relevant tags in a low-dimensional space. Hyperparameters for all methods are selected using the validation set.

Results. Table 4 presents the comparison results among TagProp, WARP, FastTag, WSABIE, CCA, and our Fast0Tag models, implemented with both linear mapping and a nonlinear neural network. TagProp significantly outperforms WARP and FastTag, but its training and testing complexities are high, at $O(M^2)$ and $O(M)$ respectively, relative to the training set size M. In contrast, WARP and FastTag are more efficient, with O(M) training complexity and constant testing complexity due to their parametric nature. Our Fast0Tag with linear mapping yields results comparable to TagProp, while Fast0Tag with the neural network surpasses the other methods. Both implementations maintain low computational complexities similar to WARP and FastTag.

Table 1: Comparison results of the conventional image tagging with 81 tags on NUS-WIDE.

Method	MiAP	K = 3			K = 5		
		P	R	F1	P	R	F1
CCA	19	9	15	11	7	20	11
WSABIE	28	16	27	20	12	35	18
TagProp	53	29	50	37	22	62	32
WARP	48	27	45	34	20	57	30
FastTag	41	23	39	29	19	54	28
Fast0Tag (lin.)	52	29	50	37	21	60	31
Fast0Tag (net.)	55	31	52	39	23	65	34

5.3 Zero-Shot and Seen/Unseen Image Tagging

This section presents results for two novel image tagging scenarios: zero-shot and seen/unseen tagging.

Fu et al. formalised the zero-shot image tagging problem, which aims to annotate test images using a pre-defined set U of unseen tags. Our Fast0Tag naturally applies to this scenario by simply ranking the unseen tags with equation (3). Furthermore, this paper also considers seen/unseen image tagging,

which finds both relevant seen tags from S and relevant unseen tags from U for the test images. The set of unseen tags U could be open and dynamically growing.

In our experiments, we treat the 81 concepts with high-quality user annotations in NUS-WIDE as the unseen set U for evaluation and comparison. We use the remaining 925 out of the 1000 frequent Flickr tags to form the seen set S - 75 tags are shared by the original 81 and 1,000 tags.

Baselines. Our Fast0Tag models can be readily applied to the zero-shot and seen/unseen image tagging scenarios. For comparison, we study the following baselines.

Seen2Unseen. We first propose a simple method that extends an arbitrary traditional image tagging method to also work with previously unseen tags. It originates from our analysis experiment in Section 3. First, we use any existing method to rank the seen tags for a test image. Second, we train a ranking SVM in the word vector space using the ranking list of the seen tags. Third, we rank unseen (and seen) tags using the learned SVM for zero-shot (and seen/unseen) tagging.

LabelEM. The label embedding method achieves impressive results on zero-shot classification for fine-grained object recognition. If we consider each tag of S U as a unique class, though this implies that some classes will have duplicated images, the LabelEM can be directly applied to the two new tagging scenarios. LabelEM+. We also modify the objective loss function of LabelEM when we train the model, by carefully removing the terms that involve duplicated images. This slightly improves the performance of LabelEM. ConSE. Again by considering each tag as a class, we include a recent zero-shot classification method, ConSE in the following experiments. Note that it is computationally infeasible to compare with Fu et al., which might be the first work to our knowledge on expanding image tagging to handle unseen tags, because it considers all the possible combinations of the unseen tags. Results. Table 5 summarizes the results of the baselines and Fast0Tag when they are applied to the zero-shot and seen/unseen image tagging tasks. Overall, Fast0Tag, with either linear or neural network mapping, performs the best.

Additionally, in the table, we add two special rows whose results are mainly for reference. The Random row corresponds to the case when we return a random list of tags in U for zero-shot tagging (and in U S for seen/unseen tagging) to each test image. We compare this row with the row of Seen2Unseen, in which we extend TagProp to handle the unseen tags. We can see that the results of Seen2Unseen are significantly better than randomly ranking the tags. This tells us that the simple Seen2Unseen is effective in expanding the labeling space of traditional image tagging methods. Some tag completion methods may also be employed for the same purpose as Seen2Unseen. Another special row in Table 5 is the last one with RankSVM for zero-shot image tagging. We obtain its results through the following steps. Given a test image, we assume the annotation of the seen tags, S, are known and then learn a ranking SVM with the default regularization = 1. The learned SVM is then used to rank the unseen tags for this image. One may wonder that the results of this row should thus be the upper bound of our Fast0Tag implemented based on linear regression because the ranking SVM models are the targets of the linear regression. However, the results show that they are not. This is not surprising, but rather it reinforces our previous statement that the learned ranking SVMs are not the "true" principal directions. The Fast0Tag implemented by the neural network is an effective alternative for seeking the principal directions. It would also be interesting to compare the results in Table 5 (zero-shot image tagging) with those in Table 4 (conventional tagging), because the experiments for the two tables share the same testing images and the same candidate tags; they only differ in which tags are used for training. We can see that the Fast0Tag (net.) results of the zero-shot tagging in Table 5 are actually comparable to the conventional tagging results in Table 4, particularly about the same as FastTag's. These results are encouraging, indicating that it is unnecessary to use all the candidate tags for training in order to have high-quality tagging performance. Annotating images with 4,093 unseen tags. What happens when we have a large number of unseen tags showing up at the test stage? NUS-WIDE provides noisy annotations for the images with over 5,000 Flickr tags. Excluding the 925 seen tags that are used to train models, there are 4,093 remaining unseen tags. We use the Fast0Tag models to rank all the unseen tags for the test images, and the results are shown in Table 3. Noting that the noisy annotations weaken the credibility of the evaluation process, the results are reasonably low but significantly higher than the random lists. Qualitative results. Figure 6 shows the top five tags for some exemplar images, returned by Fast0Tag under the conventional, zero-shot, and seen/unseen image tagging scenarios. Those by TagProp under the conventional tagging are shown on the rightmost. The tags in green color appear in the ground truth

annotation; those in red color and italic font are the mistaken tags. Interestingly, Fast0Tag performs equally well for traditional and zero-shot tagging and makes even the same mistakes.

6 Experiments on IAPRTC-12

We present another set of experiments conducted on the widely used IAPRTC-12 dataset. We use the same tag annotation and image training-test split as described in prior work for our experiments. There are 291 unique tags and 19,627 images in IAPRTC-12. The dataset is split into 17,341 training images and 2,286 testing images. We further separate 15

6.1 Configuration

Similar to the experiments in the previous section, we evaluate our methods in three distinct tasks: conventional tagging, zero-shot tagging, and seen/unseen tagging. Unlike NUS-WIDE, where a relatively small set of 81 tags is considered the ground truth annotation, all 291 tags of IAPRTC-12 are typically used in prior work to compare different methods. Therefore, we also use all of them for conventional tagging. For the zero-shot and seen/unseen tagging tasks, we exclude 20The visual features, evaluation metrics, word vectors, and baseline methods remain the same as described in the main text.

6.2 Results

Tables 4 and 5 display the results for all three image tagging scenarios (conventional, zero-shot, and seen/unseen tagging). The proposed Fast0Tag continues to outperform the other competitive baselines on this new IAPRTC-12 dataset. A notable observation, which is less apparent on NUS-WIDE probably due to its noisier seen tags, is the significant performance gap between LabelEM+ and LabelEM. This indicates that traditional zero-shot classification methods may not be directly suitable for either zero-shot or seen/unseen image tagging tasks. However, performance can be improved by tweaking LabelEM and carefully removing terms in its formulation that involve comparisons of identical images.

7 More Qualitative Results

In this section, we provide additional qualitative results from different tagging methods on both the NUS-WIDE and IAPRTC-12 datasets. These are presented to supplement the findings discussed in the main text. Due to the incompleteness and noise in tag ground truth, many accurate tag predictions are often incorrectly assessed as mistakes because they don't match the ground truth. This issue is particularly evident in the 4k zero-shot tagging results, where a wide variety of tag candidates are considered.

8 Conclusion

We have conducted a thorough examination of a specific visual pattern in words: the visual association rule that divides words into two distinct groups based on their relevance to an image. We also investigated how this rule is captured by vector offsets within the word vector space. Our empirical findings demonstrate that for any given image, there exists a main direction in the word vector space along which vectors of relevant tags are ranked higher than those of irrelevant tags. While our experimental analyses involved 1,006 words, future research should encompass larger-scale and theoretical investigations. Based on this discovery, we developed a Fast0Tag model to address image tagging by estimating the primary directions for input images. Our method is as efficient as FastTag and is capable of annotating images with a large number of previously unseen tags. Extensive experiments confirm the effectiveness of our Fast0Tag approach.