Subject2Vec: Generative-Discriminative Approach from a Set of Image Patches to a Vector

Sumedha Singla[1], Mingming Gong[2], Siamak Ravanbakhsh[3], Frank Sciurba[4], Barnabas Poczos[5], and Kayhan N. Batmanghelich[1,2,5(✉)]

[1] Computer Science Department, University of Pittsburgh, Pittsburgh, PA, USA
kayhan@pitt.edu
[2] Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA, USA
[3] Computer Science Department, University of British Columbia, Vancouver, Canada
[4] University of Pittsburgh School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA
[5] Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract. We propose an attention-based method that aggregates local image features to a subject-level representation for predicting disease severity. In contrast to classical deep learning that requires a fixed dimensional input, our method operates on a *set* of image patches; hence it can accommodate variable length input image without image resizing. The model learns a clinically interpretable subject-level representation that is reflective of the disease severity. Our model consists of three mutually dependent modules which regulate each other: (1) a *discriminative* network that learns a fixed-length representation from local features and maps them to disease severity; (2) an *attention* mechanism that provides interpretability by focusing on the areas of the anatomy that contribute the most to the prediction task; and (3) a *generative* network that encourages the diversity of the local latent features. The generative term ensures that the attention weights are non-degenerate while maintaining the relevance of the local regions to the disease severity. We train our model end-to-end in the context of a large-scale lung CT study of Chronic Obstructive Pulmonary Disease (COPD). Our model gives state-of-the art performance in predicting clinical measures of severity for COPD. The distribution of the attention provides the regional relevance of lung tissue to the clinical measurements.

1 Introduction

We propose a deep learning model that learns subject-level representation from a *set* of local features. Our model represents the image volume as a *bag* (or set) of local features (or patches) and can accommodate input images of variable sizes. We target diseases where the pathology is diffused and is not always located in the same anatomical region. The model learns by optimizing the objective

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11070, pp. 502–510, 2018.
https://doi.org/10.1007/978-3-030-00928-1_57

function that balances two goals: (1) to build a fixed length subject-level feature that is predictive of the disease severity, (2) to extract interpretable local features that identify regions of anatomy that contribute the most to the disease. Our motivation comes from the study of COPD, but the proposed model is applicable to a wide range of heterogeneous disorders.

Many diseases such as emphysema are highly heterogeneous [15] and show diffuse pattern in computed tomographic (CT) images of the lung. Having an objective way to characterize local patterns of the disease is important in diagnosis, risk prediction, and sub-typing [4,6,12,17]. Although various intensity and texture based feature descriptors are proposed to characterize the visual appearance of the disease [1,18,20], most image features are generic and are not necessarily optimized for the disease. Recent advances in deep learning enable researchers going directly from raw image to clinical outcome without specifying radiological features [3,5]. However, the classical deep learning methods, that operate on entire volume or slices [5], are challenging to interpret and they require resizing the input images to a fixed dimension. Reshaping voxels in a CT image without adjusting for the density, changes the interpretation of the intensity values.

In this paper, we view each subject as a *set* of image patches from the lung region. Previously, [1,16] viewed the subjects as sets and used handcrafted image features. In contrast, the *discriminative* part of our model uses deep learning approach and directly extracts features from the volumetric patches. Next, we use an attention mechanism [19] to adaptively weight local features and build the subject level representation, which is predictive of the disease severity. Our model is inspired by the Deep Set [21]. We extend it by adapting *generative* regularization, which prevents the redundancy of the hidden features. Furthermore, the *attention* mechanism provides interpretability by quantifying the relevance of a region to the disease. We evaluate the performance of our method on a simulated dataset and a COPD lung CT dataset where our method gives state-of-the-art performance in predicting the clinical measurements.

2 Method

We represent each subject as a set (bag) of volumetric image patches extracted from the lung region $\mathcal{X}_i = \{x_{ij}\}_{j=1}^{N_i}$, where N_i is the number of patches for subject i, which varies with subject. Our method maps x_{ij} to a low-dimensional latent space. It then aggregates the latent features to form a fix-length representation, by adaptively weighting the patches based on their contribution in prediction of disease severity (y_i). The general idea of our approach is shown in Fig. 1.

The method consists of three networks that are trained jointly: (1) a *discriminative* network, that aggregates the local information from patches in the set $\mathcal{X}_i$ to predict the disease severity y_i, (2) an *attention* mechanism, that helps discriminative network to selectively focus on patch-features by assigning weights to the patches in $\mathcal{X}_i$, and (3) a *generative* network, that regularizes the discriminative network to avoid redundant representation of patches in the latent space. The model is trained end to end, by minimizing the below objective function:

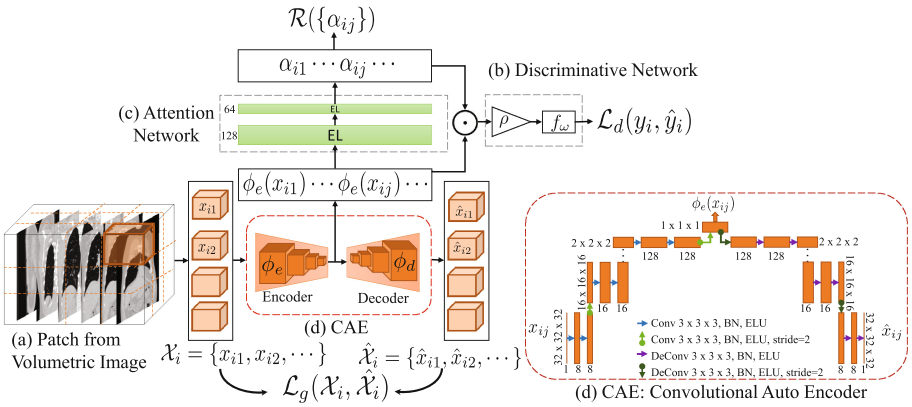

Fig. 1. (a) A subject is represented as a set of 3d image patches, (b) Discriminative Network: aggregates local features to form a fixed length representation for the subject and predicts the disease severity, (c) Attention Network: focuses attention on critical patches to provide interpretability, (d) Convolutional Auto Encoder (Generative Network): prevents redundancy of latent features.

$$\min_{\omega,\theta_e,\theta_d,\theta_a} \sum_i \mathcal{L}_d\left(y_i, \hat{y}_i(\mathcal{X}_i); \theta_e, \omega\right) + \lambda_1 \mathcal{L}_g\left(\mathcal{X}_i, \hat{\mathcal{X}}_i; \theta_e, \theta_d\right) + \lambda_2 \mathcal{R}\left(\mathcal{X}_i; \theta_e, \theta_a\right), \quad (1)$$

where $\mathcal{L}_d(\cdot,\cdot)$ and $\mathcal{L}_g(\cdot,\cdot)$ are the discriminative and generative loss functions respectively and $\mathcal{R}(\cdot)$ is a regularization over the attention. The θ_e, θ_d, θ_a and ω are the parameters of each term. λ_1, λ_2 controls the balance between the terms. The sum is over number of subjects. Next, we discuss each term in more detail.

2.1 Discriminative Network

The discriminative network transforms the input set of image patches and estimates the disease severity $\hat{y}_i(\mathcal{X}_i)$ as

$$\hat{y}_i(\mathcal{X}_i) = f\left(\rho\left(\phi_e\left(\mathcal{X}_i, \theta_e\right)\right), \omega\right). \quad (2)$$

The transformation is composed of three functions: (1) $\phi_e(\cdot;\theta_e)$ is an encoder function parameterized by θ_e. It extracts features from patches in the set $\mathcal{X}_i$ and outputs a set of features. (2) The $\rho(\cdot)$ function operates on the elements of the set and converts the variable length set $\phi_e(\mathcal{X}_i;\theta_e)$ into a fixed length vector. It is a permutation invariant function such as, maximum function $\rho(\cdot) = \max(\phi_e(x_{i,1}), \cdots, \phi_e(x_{i,N_i}))$ or mean function $\rho(\cdot) = \frac{1}{N_i}\sum_{j=1}^{N_i} \phi_e(x_{ij})$. This formulation ensures that, $\hat{y}_i(\mathcal{X}_i)$ is invariant to the order of patches in $\mathcal{X}_i$. We tried different ρ's and the mean function works well for our task. The mean function assumes all the instances within the set are contributing equally to the set-level feature vector. We extended it further to perform weighted mean, where weights are learned using the attention network in Sect. 2.2. (3) $f(\cdot;\omega)$ is a prediction function, parameterized by ω. It takes the set-level feature vector extracted

by $\rho(\cdot)$ as input, and estimates the disease severity. Finally, $\mathcal{L}_d\left(y_i, \hat{y}_i(\mathcal{X}_i); \theta_e, \omega\right)$ is a ℓ_2 loss function between predicted and true value.

2.2 Attention Mechanism

The goal of our proposed model is twofold: first to provide a prediction of the disease severity and secondly, to provide a qualitative assessment of our prediction. Here, it is reasonable to assume that different regions in the lung contribute differently to the disease severity. We model this contribution by adaptively weighting the patches. The weight indicates the importance of a patch in predicting the overall disease severity of the lung. This idea is similar to attention mechanism in Computer Vision [19] and Natural Language Processing [9] communities.

We estimate the attention weights for the subject i ($\boldsymbol{\alpha_i} = \{\alpha_{i1}, \cdots, \alpha_{i,N_i}\}$) by the attention network as

$$\boldsymbol{\alpha_i} = A\left(\phi_e\left(\mathcal{X}_i; \theta_e\right); \theta_a\right). \tag{3}$$

Unlike the $\rho(\cdot)$ function, $A(\cdot; \theta_a)$ maps a set to another set. Permuting the order of elements in the set $\mathcal{X}_i$, should *equivariantly* permute the output set $\boldsymbol{\alpha_i}$. To ensure $A(\cdot)$ is a permutation equivariant function, we construct it as a neural network with equivariant layer (EL) [21]. Assuming $\mathbf{H}_i \in \mathbb{R}^{N_i, d}$ where k^{th} row is $\phi(x_{ik}; \theta_e) \in \mathbb{R}^d$, one possible way of modeling the equivariant layer is

$$[\mathbf{H}_i]_k = \mathbf{W}\left([\mathbf{H}_i]_k - \max(\mathbf{H}_i, 1)\right) + \mathbf{b}, \tag{4}$$

where $[\mathbf{H}_i]_k$ denotes k^{th} row of $\mathbf{H}_i$ and $\max(\mathbf{H}_i, 1)$ is the max over rows. $\mathbf{W} \in \mathbb{R}^{L \times d}$, $\boldsymbol{b} \in \mathbb{R}^L$ are the parameters of the EL. To ensure $A(\cdot; \theta_a)$ is permutation equivariant we construct it by composing few EL's. Also, we assume that the weights ($\boldsymbol{\alpha_i}$) are non-negative numbers that sums to 1. The output of the EL is passed to a softmax to obtain a distribution of weights over the patches. Finally, to ensure the weights are sparsely distributed, we added a regularization term $\mathcal{R}\left(\mathcal{X}_i; \theta_e, \theta_a\right) = \sum_{j=1}^{N_i} \log(\alpha_{ij} + \epsilon)$ to the loss function in Eq. 1

2.3 Generative Network

The encoder function ϕ_e projects the raw patch x_{ij} to a d-dimensional latent representation $\left(i.e., \phi_e(x_{ij}; \theta_e) \in \mathbb{R}^d\right)$. Without extra regularization, the loss function focuses only on the prediction task, forcing ϕ_e to extract information that is only relevant to y. If y is low dimensional, ϕ_e learns a highly redundant latent space representation for each patch. Since α_{ij} is a function of $\phi_e(x_{ij}, \theta_e)$, redundant features result in uniform weights *i.e.,* $(\alpha_{ij} = \frac{1}{|\mathcal{X}_i|})$. This phenomenon makes interpretability difficult. We demonstrated this effect in our experiments.

To discourage loss of information, we added a convolutional auto-encoder (CAE) [11] to reconstruct patch as $\hat{x}_{ij} = \phi_d(\phi_e(x_{ij}; \theta_e); \theta_d)$. A generative loss $\mathcal{L}_g(\mathcal{X}_i, \hat{\mathcal{X}}_i; \theta_e, \theta_d) = \frac{1}{|\mathcal{X}_i|} \sum_{x_{ij} \in \mathcal{X}_i} ||x_{ij} - \hat{x}_{ij}||_2$ is added to the final loss function.

Table 1. Clinical measurement regression and GOLD stage classification accuracy by different methods on the COPDGene dataset.

Method	FEV1	FEV1/FVC	GOLD exact	GOLD one-off
Our method ($\lambda_1 = 0$)	**0.68**	**0.71**	**61.17** %	**87.64** %
Our method ($\lambda_1 = 10$)	**0.64**	0.70	**55.60** %	**84.57**%
CNN [5]	0.53	—	51.1 %	74.9 %
Non-Parametric [16]	0.58	0.70	50.47 %	—
K-Means [16]	0.54	0.67	48.23 %	—
Baseline	0.52	0.69	49.06 %	—

2.4 Architecture Details

The $f(\cdot;\omega)$ is a linear function predicting the disease severity y_i. The architecture of generative network is elaborated in Fig. 1. The convolutional layer employs batch-normalization for regularization, followed by an exponential linear unit (ELU) [2] for non-linearity. The attention network $A(\cdot;\theta_a)$ has 2 equivalence layers with sigmoid activation function, followed by a softmax layer. The model is trained using Adam optimizer [7] with a fixed learning rate of 0.001.

3 Experiments

We evaluate the prediction and interpretation of our method on synthetic and real datasets. To evaluate the interpretability of our method quantitatively, we synthesize a dataset where the set-level target (y) are simulated from a subset of instances. Hence by viewing the attention weights as a detector of the relevant instances, we are able to evaluate the interpretability of our approach.

3.1 Synthetic Data

In this experiment, we build 10,000 training and 8,000 testing sets. The instances in the set are randomly drawn images from MNIST [8] dataset. The size of the sets varies between 20 to 100 instances. Each image is a 28×28 pixel monochrome image of a handwritten digit between $0-9$. The set-label (y) is the sum of prime numbers $(2, 3, 5, 7)$ in that set. Our method predicts the set-label with a high accuracy ($R^2 = 0.99$ on held-out data). We view the attention weights as detectors of prime numbers. Note that no instance level supervision is used. We make an ROC (Receiver Operating Characteristic) curve per set, and compute one average ROC curve across the held-out dataset. Figure 2(a) shows the average and error bar for all the sets. The figure compares our method (blue) with equal weights (red) (*i.e.,* $\alpha_{ij} = 1/|\mathcal{X}_i|$) and uniform random weights (green). Our method can detect correct instances in the set, with only weak supervision over the set (*i.e.,* set-level label y). Here we used $\lambda_1 = 100$ and $\lambda_2 = 0.01$.

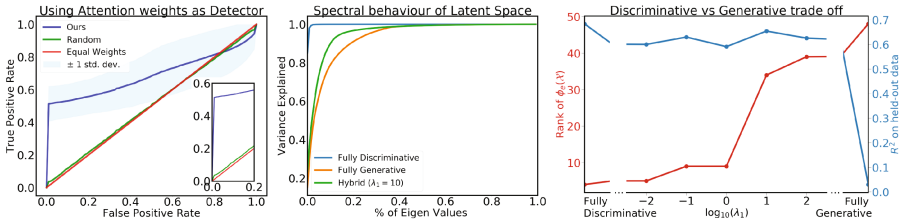

Fig. 2. (a) ROC curve of detecting true relevant instances on synthetic dataset using attention weights, (b) Spectral properties of patch-level features for different values of λ_1. (c) The trade-off between rank of latent space (red, y-axis on left) and predictive power (blue, y-axis on right) for different values of λ_1. Left represents fully discriminative and right represents fully generative models.

3.2 COPD

We evaluate our model on 6,400 subjects with different degrees of severity of the COPD from the COPDGene dataset [13]. As clinical measures, we use the Forced Expiratory Volume in one second (FEV1), the ratio of FEV1 and Forced Vital Capacity (FVC), and discrete score (between 0–4) called the Global Initiative for Chronic Obstructive Lung Disease (GOLD). We first segment the lung area on the inspiratory images using CIP library [14]. Each subject is represented as a bag of equal size 3D patches, with some overlap. Large patch size and percentage overlap leads to GPU memory issues. We experimented with different values and finally used patch-size of $32 \times 32 \times 32$ with 40% overlap in our experiments.

We perform three experiments: (1) *Prediction:* we compare the performance of our method against the sate-of-art for predicting the clinical measurements, (2) *Generative regularizer (λ_1):* we study the effect of the generative regularizer (*i.e.,* λ_1) in terms of prediction accuracy and information preserved in latent space, (3) *Visualization:* we visualize the interpretation of the model on the subject and population level. Unlike λ_1, the choice of λ_2 don't have any significant effect on the prediction accuracy. The value of λ_2 influences the sparsity and diversity of the attention weights. In the experiments, we fixed λ_2 to 0.0001.

Prediction: We compare to several baselines: (a) **Baseline**: Two threshold-based features measuring the percentage of voxels with intensity less than -950 Hounsfield Unit (HU) for the inspiratory and -856 HU for expiratory images. These measurements reflect the clinical measure to quantify emphysema and the degree of gas trapping. (b) **Non-parametric**: Schabdach et al. [16] view each subject as a set of hand-crafted histogram and texture features from supervoxels. They represent each subject in an embedding space using a non-parametric distance between sets. (c) **CNN**: Gonzalez et al. [5] use deep features learned from a composite image of four canonical views of a CT scan to quantify FEV1 and stage COPD. (d) **BOW**: It extracts features similar to [16] from supervoxels, but applies $k-$means to extract the subject-level representation. We perform

Fig. 3. (a) Embedding the subjects in 2D using tSNE. The dots represents one subject colored by the GOLD score. (b) An axial view of the attention map on a subject. Red color indicate higher relevance to the disease severity.

10-fold cross-validation and report R^2 for the continuous measurements (*i.e.,* FEV1 and FEV/FVC) and accuracy for the GOLD score. For GOLD score we also report the percentage of cases whose classification lays within one class of the true value (*one-off*). The Table 1 summarizes the results of the experiments. Our method outperforms the state-of-the-art on predicting FEV1 and GOLD score. Adding the generative regularization ($\lambda_1 = 10$) reduces the accuracy but provides better interpretability. In the following, we study the effect of λ_1.

Generative regularizer (λ_1): The Fig. 2(b) reports the spectral behaviors of the latent features (*i.e.,* $\phi_e(\mathcal{X}_i)$) for varying λ_1. For small λ_1 the loss function doesn't optimize for the generative loss. Hence, the latent space representation becomes highly redundant, and all the attention weights α_{ij} converges to $\frac{1}{|\mathcal{X}_i|}$. The Fig. 2(c) shows the trade-off between effective rank of the latent feature and R^2 for predicting FEV1. Although, the R^2 drops a little, the rank, which represents the diversity of the latent features, improves drastically. The gap between accuracies of $\lambda_1 = 0$ and $\lambda_1 > 0$ is the price we pay for the interpretability. Fully generative model ($\lambda_1 \to \infty$) does not produce good prediction.

Visualization: We use tSNE [10] to visualize subject-level features in two dimension. In Fig. 3(a), each dot represents a subject colored by the GOLD score. Even in two dimension, subjects with GOLD score of (0,1) and (3,4) are quite separable and 2's are in between. The bimodal distribution of GOLD stages 3 and 4, is sensitive to t-SNE parameterization and requires further investigation. 3(b) visualizes the attention weights on one subject. The dark area on the left lung, which is severely damaged, received hight attention.

4 Conclusion

We developed a novel attention-based model that achieves high prediction while maintaining interpretability. The method outperforms state-of-art and detects correct instances on the simulated data. Our current model does not account

for spatial locations of the patches. As a future direction, we plan to extend the model to accommodate relationship between patches.

Acknowledgement. This work is partially supported by NIH Award Number 1R01HL141813-01. We gratefully thank NVIDIA Corporation for their donation of the Titan X Pascal GPU. We thank Competitive Medical Research Fund (CMRF) grant for their funding.

References

1. Cheplygina, V., Peña, I.P., Pedersen, J.H., Lynch, D.A., Sørensen, L., de Bruijne, M.: Transfer learning for multi-center classification of chronic obstructive pulmonary disease, January 2017
2. Clevert, D.A., Unterthiner, T., Hochreiter, S.: Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), November 2015
3. Dubost, F., et al.: GP-Unet: lesion detection from weak labels with a 3D regression network. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 214–221. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_25
4. Estépar, R.S.J., Kinney, G.L.: Computed tomographic measures of pulmonary vascular morphology in smokers and their clinical implications. AJRCCM **188**(2), 231–239 (2013)
5. González, G., Ash, S.Y., Vegas-Sánchez-Ferrero, G.: Disease staging and prognosis in smokers using deep learning in chest computed tomography. AJRCCM **197**(2), 193–203 (2017)
6. Hayhurst, M.D., MacNee, W., Flenley, D.C.: Diagnosis of pulmonary emphysema by computerised tomography. Lancet **2**(8398), 320–322 (1984)
7. Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization, December 2014
8. LeCun, Y., Cortes, C.: MNIST handwritten digit database. AT&T Labs (2010)
9. Luong, M.T., Pham, H., Manning, C.D.: Effective Approaches to Attention-based Neural Machine Translation, August 2015
10. van der Maaten, L., Hinton, G., Visualizing Data using t-SNE: Visualizing Data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
11. Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: Honkela, T., Duch, W., Girolami, M., Kaski, S. (eds.) ICANN 2011. LNCS, vol. 6791, pp. 52–59. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-21735-7_7
12. Müller, N.L., Staples, C.A., Miller, R.R., Abboud, R.T.: Density mask: an objective method to quantitate emphysema using computed tomography. Chest **94**, 782–787 (1988)
13. Regan, E.A., et al.: Genetic epidemiology of COPD (COPDGene) study design. J. COPD **7**(1), 32–43 (2010)
14. San Jose Estepar, R., Ross, J.C., Harmouche, R., Onieva, J., Diaz, A.A., Washko, G.R.: CIP: an open-source library and workstation for quantitative chest imaging. Am. J. Respir. Crit. Care Med. **191**, A4975 (2015)
15. Satoh, K., Kobayashi, T., Murota, M.: CT assessment of subtypes in pulmonary emphysema in smokers. JJCR **46**(1), 98–102 (2001)

16. Schabdach, J., Wells, W.M., Cho, M., Batmanghelich, K.N.: A likelihood-free approach for characterizing heterogeneous diseases in large-scale studies. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 170–183. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_14
17. Shapiro, S.D.: Evolving concepts in the pathogenesis of chronic obstructive pulmonary disease. Clin. Chest Med. **21**(4), 621–632 (2000)
18. Sorensen, L., Nielsen, M., Lo, P., Ashraf, H., Pedersen, J.H., de Bruijne, M.: Texture-based analysis of COPD: a data-driven approach. IEEE Trans. Med. Imaging **31**(1), 70–78 (2012)
19. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, February 2015
20. Yang, J., et al.: Unsupervised discovery of spatially-informed lung texture patterns for pulmonary emphysema: the MESA COPD study. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10433, pp. 116–124. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66182-7_14
21. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R., Smola, A.: Deep sets. In: Advances in Neural Information Processing Systems, pp. 3391–3401, March 2017