A New Method for Structured Learning with Privileged Information

Shiding Sun[1], Chunhua Zhang[1] *, and Yingjie Tian[2]

[1] School of Information, Renmin University of China, Beijing 100872, China
[2] Research Center on Fictitious Economy and Data Science, Chinese Academy of Science, Beijing 100190, China

Abstract. In this paper, we present a new method JKSE+ for structured learning. Compared with some classical methods such as SSVM and CRFs, the optimization problem in JKSE+ is a convex quadratical problem and can be easily solved because it is based on JKSE. By incorporating the privileged information into JKSE, the performance of JKSE+ is improved. We apply JKSE+ to the problem of object detection, which is a typical one in structured learning. Some experimental results show that JKSE+ performs better than JKSE.

Keywords: SVM, One-Class SVM, Structured Learning, Object Detection, Privileged Information

1 Introduction

This paper deals with the structured learning problems which learn function: $f : \mathcal{X} \to \mathcal{Y}$, where the elements of $\mathcal{X}$ and $\mathcal{Y}$ are structured objects such as sequences, trees, bounding boxes, strings. Structured learning arises in lots of real world applications including multi-label classification, natural language parsing, object detection, and so on. Conditional random fields [5, 6], maximum margin markov networks [9] and structured output support vector machines(SSVM) [10] have been developed as powerful tools to predict the structured data. The common approach of these methods is to define a linear scoring function based on a joint feature map over inputs and outputs. There are some drawbacks in these methods. On the one hand, to apply them one requires clearly labeled training sets. Experiments show that some incorrect or incomplete labels can reduce their performance. On the other hand, training these models is computationally cost. So it is difficult or infeasible to solve large scale problems except for some special output structures.

To overcome these drawbacks, a method called Joint Kernel Support Estimation(JKSE) has been proposed in [7]. JKSE is a generative method as it relies

* Corresponding author. Email: zhangchunhua@ruc.edu.cn. This work has been partially supported by grants from National Natural Science Foundation of China (Nos. 61472390, 71731009, 71331005, 91546201 and 11771038), and the Beijing Natural Science Foundation (No.1162005).

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_35

on learning the support of the joint-probability density of inputs and outputs. This makes it robust in handling mislabeled data. At the same time, The optimization problem is convex and can be efficiently solved because the one-class SVM is used in it. However, JKSE is not as powerful as SSVM [2]. So we focus on the following problem: How to improve the performance of JKSE? To answer this question, we introduce the privileged information into JKSE.

Privileged information [11] provides useful high-level knowledge that is used only at training time. For example, in the problem of object detection, these information includes the object's parts, attributes and segmentations. More reliable models [11, 4, 3, 8] can be learned by incorporating these high-level information into SVM, SSVM, one-class SVM.

In this paper, we propose a new method called JKSE+ based on JKSE with privileged information and apply it to the problem of object detection. Some experiments show that our new method JKSE+ performs better than JKSE.

The rest of this paper is organized as follows. We first review the method JKSE in section 2 , then introduce our new method JKSE+ in section 3, and the experimental results are presented in section 4.

2 Related work

This section considers the following structured learning problem: given the training set: $\{(x_1,y_1),...,(x_l,y_l)\}$, where $x_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$. $\mathcal{X}$ and $\mathcal{Y}$ are the space of inputs and outputs with some structures respectively. Assume that the input-output pairs (x,y) follow a joint probability distribution $p(x,y)$. Our goal is to learn a mapping: $g : \mathcal{X} \to \mathcal{Y}$ such that for a new input $x \in \mathcal{X}$, the corresponding label $y \in \mathcal{Y}$ can be determined by maximizes the posterior probability $p(y|x)$.

As we all know, The discriminative method directly models the conditional distribution $p(y|x)$, and the generative method directly models the joint distribution $p(x,y)$. These two methods are equivalent, i.e. $\arg\max_{y\in\mathcal{Y}} p(y|x) = \arg\max_{y\in\mathcal{Y}} p(x,y)\ for\ any\ x \in \mathcal{X}$. JKSE is a generative method. Suppose that $p(x,y) = \frac{1}{Z}\exp(\langle w, \Phi(x,y)\rangle)$. Here, $Z \equiv \sum_{x,y}\exp(\langle w, \Phi(x,y)\rangle)$, and Z is a normalization constant. We can ignore Z during training and testing. The JKSE method translates the task of learning a joint probability distribution $p(x,y)$ into a one-class SVM problem to estimate the joint probability distribution $p(x,y)$.

In training phase, JKSE solves the following problem:

$$\begin{aligned} &\min_{w,\xi,\rho} \tfrac{1}{2}\parallel w\parallel^2 + \tfrac{1}{vl}\sum_{i=1}^{l}\xi_i - \rho \\ &s.t.\ \langle w, \Phi(x_i,y_i)\rangle \geq \rho - \xi_i, \quad i = 1,2,...,l, \\ &\quad \xi_i \geq 0, \quad i = 1,2,...,l. \end{aligned} \tag{1}$$

To get its solution, JKSE solve its dual problem:

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_35

$$\begin{aligned} \min_{\alpha} & \sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K\left(\left(x_i,y_i\right),\left(x_j,y_j\right)\right) \\ s.t. \quad & 0\leq\alpha_i\leq\tfrac{1}{vl}, \quad i=1,...,l, \\ & \sum_{i=1}^{l}\alpha_i=1. \end{aligned} \tag{2}$$

where $K\left(\left(x,y\right),\left(x',y'\right)\right)\equiv\left\langle\Phi\left(x,y\right),\Phi\left(x',y'\right)\right\rangle$ is a joint feature kernel function. If α^* is the solution to the above problem (2), then the solution to the primal problem (1) for w is given as follows:

$$w^*=\sum_{i=1}^{l}\alpha_i^*\Phi\left(x_i,y_i\right). \tag{3}$$

Furthermore, in the inference step, for a new input $x\in\mathcal{X}$, the corresponding label y is given by:

$$y=\arg\max_{y\in\mathcal{Y}}\sum_{i=1}^{l}\alpha_i K\left(\left(x_i,y_i\right),\left(x,y\right)\right). \tag{4}$$

3 JKSE+

Assume that we have some privileged information, $\left(x_1^*,x_2^*,...,x_l^*\right)\in\mathcal{X}^*$ that is available only at the training phase but not available on the test phase. Now we consider the following privileged structured learning problem:

Given a training set $T=\left\{\left(x_1,x_1^*,y_1\right),...,\left(x_l,x_l^*,y_l\right)\right\}$ where $x_i\in\mathcal{X}$, $x_i^*\in\mathcal{X}^*$, $y\in\mathcal{Y}$, $i=1,..,l$, our goal is to find a mapping: $g:x\rightarrow y$, such that the label of y for any x can be predicted by $y=g\left(x\right)$.

Now we discuss how the privileged information can be incorporated into the framework of JKSE. Suppose that there exists the best but unknown function: $\arg\max_{y\in\mathcal{Y}}\left\langle w_0,\Phi\left(x,y\right)\right\rangle$. The function $\xi\left(x\right)$ of the input x is defined as follows:

$$\xi^0=\xi\left(x\right)=\left[\rho-\left\langle w_0,\Phi\left(x,y\right)\right\rangle\right]_+$$

where $\left[\eta\right]_+=\begin{cases}\eta, & if \quad \eta\geq 0,\\ 0, & otherwise.\end{cases}$ If we know the value of the function $\xi\left(x\right)$ on each input $x_i\left(i=1,...,l\right)$ such as we know the triplets $\left(x_i,\xi_i^0,y_i\right)$ with $\xi_i^0=\xi\left(x_i\right),i=1,...,l$, we can get improved prediction. However, in reality, this is impossible. Instead we use a correcting function to approximate the function $\xi\left(x\right)$. Similar to one-class SVM with privileged information in [3], we replace ξ_i by a mixture of values of the correcting function $\psi\left(x_i^*\right)=\left\langle w^*,\Phi\left(x_i^*,y_i\right)\right\rangle+b^*$ and some values ζ_i, and get the primal problem of JKSE+:

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_35

$$\begin{aligned} &\min_{w,w^*,b^*,\rho,\zeta} \frac{vl}{2} \parallel w\parallel^2 + \frac{\gamma}{2} \parallel w^*\parallel^2 - vl\rho + \sum_{i=1}^{l} \left[\langle w^*, \Phi^* \left(x_i, y_i\right)\rangle + b^* + \zeta_i\right] \\ &s.t. \quad \langle w, \Phi \left(x_i, y_i\right)\rangle \geq \rho - \left(\langle w^*, \Phi^* \left(x_i^*, y_i\right)\rangle + b^*\right), \quad i = 1, ..., l, \\ &\qquad \langle w^*, \Phi^* \left(x_i^*, y_i\right)\rangle + b^* + \zeta_i \geq 0, \zeta_i \geq 0, \quad i = 1, .., l. \end{aligned} \tag{5}$$

The Lagrange function for this problem is:

$$\begin{aligned} &L\left(w, w^*, b^*, \rho, \zeta, \mu, \alpha, \beta\right) = \frac{vl}{2} \parallel w\parallel^2 + \frac{\gamma}{2} \parallel w^*\parallel^2 - vl\rho + \sum_{i=1}^{l} \left[\langle w^*, \Phi^* \left(x_i, y_i\right)\rangle + b^* + \zeta_i\right] \\ &- \sum_{i=1}^{l} \mu_i \zeta_i - \sum_{i=1}^{l} \alpha_i \left[\langle w, \Phi \left(x_i, y_i\right)\rangle - \rho + \langle w^*, \Phi^* \left(x_i^*, y_i\right)\rangle + b^*\right] \\ &- \sum_{i=1}^{l} \beta_i \left[\langle w^*, \Phi^* \left(x_i^*, y_i\right)\rangle + b^* + \zeta_i\right] \end{aligned} \tag{6}$$

The KKT conditions are as follows:

$$\nabla_w L = vlw - \sum_{i=1}^{l} \alpha_i \Phi \left(x_i, y_i\right) = 0, \tag{7}$$

$$\nabla_{w^*} L = \gamma w^* + \sum_{i=1}^{l} \Phi^* \left(x_i^*, y_i\right) - \sum_{i=1}^{l} \alpha_i \Phi^* \left(x_i^*, y_i\right) - \sum_{i=1}^{l} \beta_i \Phi^* \left(x_i^*, y_i\right), \tag{8}$$

$$\frac{\partial L}{\partial b^*} = l - \sum_{i=1}^{l} \alpha_i - \sum_{i=1}^{l} \beta_i = 0, \tag{9}$$

$$\frac{\partial L}{\partial \rho} = -vl + \sum_{i=1}^{l} \alpha_i = 0, \tag{10}$$

$$\frac{\partial L}{\partial \zeta_i} = 1 - \beta_i - \mu_i = 0, i = 1, ..., l, \tag{11}$$

$$\rho - \left(\langle w^*, \Phi^* \left(x_i^*, y_i\right)\rangle + b^*\right) - \langle w, \Phi \left(x_i, y_i\right)\rangle \leq 0, i = 1, ..., l, \tag{12}$$

$$- \left(\langle w^*, \Phi^* \left(x_i^*, y_i\right)\rangle + b^* + \zeta_i\right) \leq 0, i = 1, ..., l, \tag{13}$$

$$-\zeta_i \leq 0, i = 1, ..., l, \tag{14}$$

$$\alpha_i \left[\rho - \left(\langle w^*, \Phi^* \left(x_i^*, y_i\right)\rangle + b^*\right) - \langle w, \Phi \left(x_i, y_i\right)\rangle\right] = 0, i = 1, ..., l, \tag{15}$$

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_35

$$\beta_i \left[\langle w^*, \Phi^* \left(x_i^*, y_i\right)\rangle + b^* + \zeta_i\right] = 0, i = 1, ..., l, \tag{16}$$

$$\mu_i \zeta_i = 0, i = 1, ..., l, \tag{17}$$

$$\alpha_i \geq 0, \beta_i \geq 0, \mu_i \geq 0, i = 1, ..., l. \tag{18}$$

From the above KKT conditions and setting $\delta_i = 1 - \beta_i$, we can get that

$$w = \frac{1}{vl} \sum_{i=1}^{l} \alpha_i \Phi \left(x_i, y_i\right), \tag{19}$$

$$w^* = \frac{1}{\gamma} \sum_{i=1}^{l} \left(\alpha_i - \delta_i\right) \Phi^* \left(x_i^*, y_i\right), \tag{20}$$

$$\sum_{i=1}^{l} \delta_i = \sum_{i=1}^{l} \alpha_i = vl, \tag{21}$$

$$0 \leq \delta_i \leq 1, i = 1, ..., l. \tag{22}$$

So, we can get the dual problem is as follows:

$$\begin{aligned} \max_{\alpha,\delta} & -\tfrac{1}{2vl} \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j K\left(\left(x_i, y_i\right), \left(x_j, y_j\right)\right) \\ & - \sum_{i=1}^{l} \sum_{j=1}^{l} \tfrac{1}{2\gamma} \left(\alpha_i - \delta_i\right) K^* \left(\left(x_i^*, y_i\right), \left(x_j^*, y_j\right)\right) \left(\alpha_j - \delta_j\right) \\ s.t. \quad & \sum_{i=1}^{l} \alpha_i = vl, \quad \alpha_i \geq 0, \\ & \sum_{i=1}^{l} \delta_i = vl, \quad 0 \leq \delta_i \leq 1. \end{aligned} \tag{23}$$

We use $K\left(\left(x_i, y_i\right), \left(x_j, y_j\right)\right)$ and $K^* \left(\left(x_i^*, y_i\right), \left(x_j^*, y_j\right)\right)$ to replace the inner product $\left\langle \Phi \left(x_i, y_i\right), \Phi \left(x_j, y_j\right)\right\rangle$ and $\left\langle \Phi^* \left(x_i^*, y_i\right), \Phi^* \left(x_j^*, y_j\right)\right\rangle$. Therefore, the model's decision function is $f\left(x, y\right) = \sum_{i=1}^{l} \alpha_i K\left(\left(x_i, y_i\right), \left(x, y\right)\right)$.

We can learn this mapping in JKSE framework as

$$y = g\left(x\right) = \arg\max_{y \in \mathcal{Y}} f\left(x, y\right) = \arg\max_{y \in \mathcal{Y}} \sum_{i=1}^{l} \alpha_i K\left(\left(x_i, y_i\right), \left(x, y\right)\right). \tag{24}$$

Here, the function $f\left(x, y\right)$ is equivalent to a matching function. For example in object detection, when the overlap of an object and a bounding box is higher,

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_35

the value of the function is greater. Therefore, we output y that maximizes the value of $f(x, y)$.

Our new algorithm JKSE+ is given as follows:

Algorithm 1

(1) Given a training set $T = \{(x_1, x_1^*, y_1), ..., (x_l, x_l^*, y_l)\}$ where $x_i \in \mathcal{X}$, $x_i^* \in \mathcal{X}^*$, $y \in \mathcal{Y}$, $i = 1, .., l$;

(2) Choose the appropriate kernel function $K(u, v)$, $K^*(u', v')$ and penalty parameters $v > 0, \gamma > 0$;

(3) Construct and solve convex quadratic programming problem:

$$\begin{aligned}
\max_{\alpha,\delta} & -\tfrac{1}{2vl}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j K((x_i, y_i), (x_j, y_j)) \\
& -\sum_{i=1}^{l}\sum_{j=1}^{l}\tfrac{1}{2\gamma}(\alpha_i - \delta_i)K^*\left((x_i^*, y_i), \left(x_j^*, y_j\right)\right)(\alpha_j - \delta_j) \\
s.t. \quad & \sum_{i=1}^{l}\alpha_i = vl, \quad \alpha_i \geq 0, \\
& \sum_{i=1}^{l}\delta_i = vl, \quad 0 \leq \delta_i \leq 1.
\end{aligned}$$

get the solution $(\alpha^*, \delta^*) = (\alpha_1^*, ...\alpha_l^*, \delta_1^*, ..., \delta_l^*)$.

(4) Construct decision function:

$$y = g(x) = \arg\max_{y\in\mathcal{Y}} f(x, y) = \arg\max_{y\in\mathcal{Y}} \sum_{i=1}^{l}\alpha_i^* K((x_i, y_i), (x, y)).$$

4 Experiments

In this section, we apply our new method to the problem of object detection. In object detection, given a set of pictures, we hope to learn a mapping $g : \mathcal{X} \to \mathcal{Y}$, when inputing a picture, we can get the object's position in the picture by mapping g. Obviously, it is a typical one of structured learning and can be solved by our new method. Some experiments are made in this section.

4.1 Dataset

We use dataset Caltech-UCSD Birds 2011 (CUB-2011) [12] to evaluate our algorithm. This dataset contains two hundred species of birds, each of which has sixty pictures. Each picture contains only one bird, the bird's position in

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_35

the picture is indicated by a bounding box. In addition, this dataset provides privilege information, including the bird's attribute information for each image described as a 312-dimensional vector and segmentation masks.

4.2 Features and Privileged Information

Our feature descriptor adopts the bag-of-visual-words model based on SURF descriptor [1]. We use attribute informations and segmentation masks as privileged information. For the feature extraction of segmentation mask, we use the same strategy as the original image for feature extraction, that is SURF based bag-of-visual-words feature descriptor. It is clear that the feature space of privileged information provides more information relative to the feature space of the original image so that the object's location in the image can be better detected.

We select 50 pictures as the training set and 10 pictures as the test set. The dimensionality of original visual feature descriptors is 200. In addition, attribute information is described as a 312-dimensional vector, each dimension is a binary variable. We extract the 500-dimensional feature descriptors based on the same bag-of-visual-words model from segmentation masks as in the original picture. So the privilege information has a dimension of 812-dimensional vectors.

In Fig.1, we can see that more feature descriptors can be extracted in the segmentation masks, which is beneficial to improve the overlap of object detection.

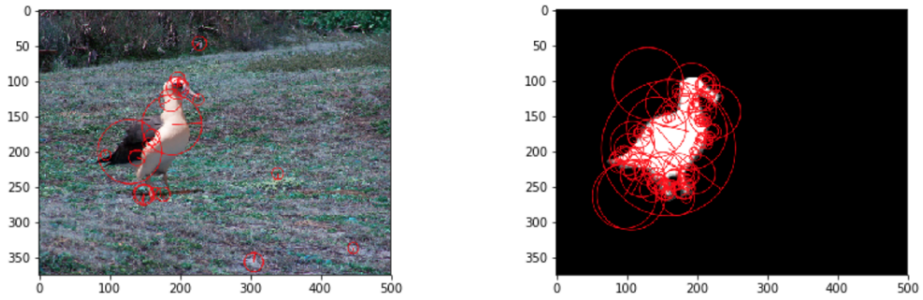

Fig. 1. The picture on the left is the feature descriptor of the original picture. The picture on the right is the feature descriptor of the segmentation mask, which is used as privilege information when training.

4.3 Kernal Function

We use the following version of the chi-square kernel function $\left(\chi^2 - \text{kernel}\right)$:

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_35

$$K(u,v) = K^*(u,v) = e^{-\theta \sum_{i=1}^{n} \frac{(u_i - v_i)^2}{u_i + v_i}}, u \in R^n, v \in R^n.$$

This kernel is most commonly applied to histograms generated by bag-of-visual-words model in computer vision [13].

4.4 Experimental results

To evaluate our JKSE+, we compare it with JKSE. During the training, we adjust the parameters v, γ, θ on a 8×8×8 space spanning values $\left[10^{-4}, 10^{-3}, ..., 10^{3}\right]$. For JKSE, we also adjust the parameter v, θ on a 8×8 space spanning values $\left[10^{-4}, 10^{-3}, ..., 10^{3}\right]$.

We chose ten different birds to compare the detection results of JKSE and JKSE +.

Table 1. Dataset

Data_ID	Name
001	Black_footed_Albatross
002	Laysan_Albatross
003	Sooty_Albatross
004	Groove_billed_Ani
005	Crested_Auklet
006	Least_Auklet
007	Parakeet_Auklet
008	Rhinoceros_Auklet
009	Brewer_Blackbird
010	Red_winged_Blackbird

Table 2. Overlap ratio of Object Detection

Model \ Data_ID	001	002	003	004	005	006	007	008	009	010
JKSE	40.974	34.281	**55.808**	28.948	38.719	47.705	51.414	31.695	**54.044**	34.285
JKSE+	**46.241**	**42.933**	46.347	**30.323**	**44.660**	**51.455**	**53.692**	**40.342**	49.919	**37.866**
DIFF.	+5.267	+8.652	-9.461	+1.375	+5.941	+3.750	+2.278	+8.647	-4.125	+3.581

The overlap ratio of JKSE+ is higher than that of JKSE in eight datasets.

5 Conclusion

We propose a new method for structured learning with privilege information based on JKSE. Firstly, compared with some traditional methods SSVM, CRFs

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_35

for structured learning, the resulting optimization problem in our new model JKSE+ is convex and can be easily solved. Secondly, compared with JKSE, the prediction performance of JKSE is improved by using the privileged information. Lastly, we apply JKSE+ to the problem of object detection. Some experimental results show that JKSE+ performs better than JKSE in most cases.

For future work, we will consider some extensions of the JKSE+ method. For example, at the training stage privileged information are provided only for a fraction of inputs or privileged information are described in many different spaces, and so on.

References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding **110**(3), 346–359 (2008)
2. Blaschko, M.B., Lampert, C.H.: Learning to localize objects with structured output regression. In: European conference on computer vision. pp. 2–15. Springer (2008)
3. Burnaev, E., Smolyakov, D.: One-class svm with privileged information and its application to malware detection. In: Data Mining Workshops (ICDMW), 2016 IEEE 16th International Conference on. pp. 273–280. IEEE (2016)
4. Feyereisl, J., Kwak, S., Son, J., Han, B.: Object localization based on structural svm using privileged information. In: Advances in Neural Information Processing Systems. pp. 208–216 (2014)
5. Lafferty, J., McCallum, A., Pereira, F.C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
6. Lafferty, J., Zhu, X., Liu, Y.: Kernel conditional random fields: representation and clique selection. In: Proceedings of the twenty-first international conference on Machine learning. p. 64. ACM (2004)
7. Lampert, C.H., Blaschko, M.B.: Structured prediction by joint kernel support estimation. Machine Learning **77**(2-3), 249 (2009)
8. Tang, J., Tian, Y., Zhang, P., Liu, X.: Multiview privileged support vector machines. IEEE transactions on neural networks and learning systems (2017)
9. Taskar, B., Guestrin, C., Koller, D.: Max-margin markov networks. In: Advances in neural information processing systems. pp. 25–32 (2004)
10. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. Journal of machine learning research **6**(Sep), 1453–1484 (2005)
11. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural networks **22**(5-6), 544–557 (2009)
12. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
13. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International journal of computer vision **73**(2), 213–238 (2007)

ICCS Camera Ready Version 2018
To cite this paper please use the final published version:
DOI: 10.1007/978-3-319-93701-4_35