Agency for Science, Technology and Research
SINGAPORE

CREATING GROWTH, ENHANCING LIVES

The 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)

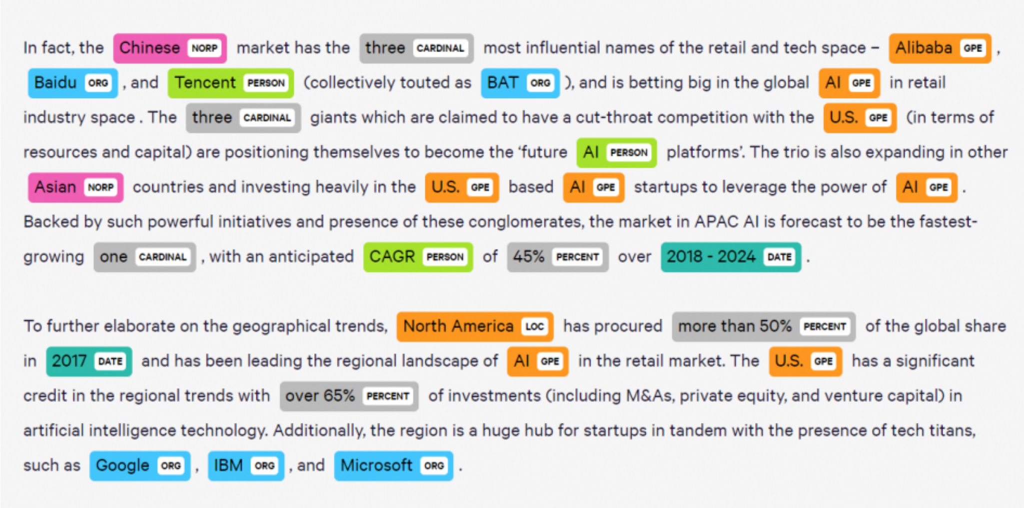
DATNet: Dual Adversarial Neural Transfer for Low-Resource Named Entity Recognition

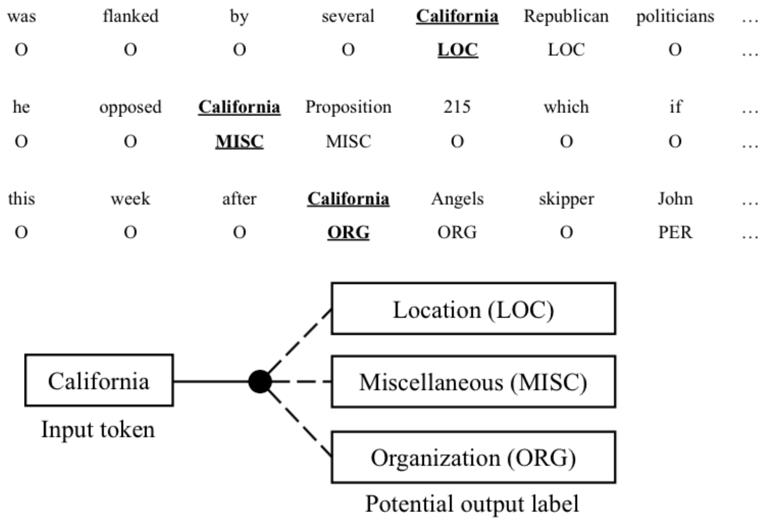
Joey Tianyi Zhou*, Hao Zhang*, Di Jin, Hongyuan Zhu, Meng Fang, Rick Siow Mong Goh and Kenneth Kwok

DATNet: Background

Named entity recognition, also known as **NER**, classifies named entities that are present in a text into pre-defined categories like person, organization, location, dates, etc.

NER is **challenging** and detects not only **the type of named entity**, but also **the entity boundaries**, which requires deep understanding of *contextual semantics* to **disambiguate** the *different entity types of same tokens*.





DATNet: Background

Traditional Method for NER

- Conditional Random Field (CRF), Support Vector Machine (SVM), Perceptron, etc.
- Hand-craft features by expertise.

➢ Drawbacks: require a lot of domain-knowledge to design features.

Deep Learning for NER

- Deep Neural Nets (DNN), Convolutional Nets (CNN), Recurrent Nets (RNN), etc.
- Requires little feature engineering and domain knowledge.

➢ Limitations: mass of data is required for better generalization ability.

Transfer Learning for Low-resource NER

- When annotated corpora is small, NN-based methods degrade significantly, since hidden features cannot be learned adequately.
- *Transfer learning* is a way to overcome such obstacle by borrowing knowledge from other resources.

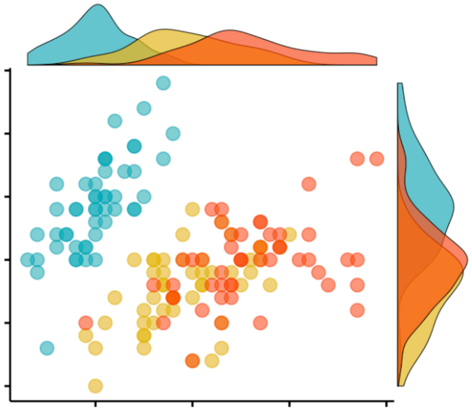
DATNet: Background

Although the existing transfer-based methods show promising performance in low-resource settings. There are two issues deserved to be further investigated on:

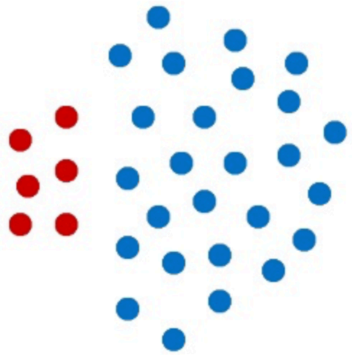
1. **Representation Difference**: They did not consider the representation difference across source and target in different scenarios (Cross-languages/domains).
2. **Resource Data Imbalance**: the training size of high-resource is usually much larger than that of low-resource.

Most existing methods ignore the above two issues in their models, thus resulting in poor generalization.

The Dual Adversarial Transfer Nets (DATNet) is proposed to solve these two issues.



Representation difference



Resource Data Imbalance

Agency for Science, Technology and Research
SINGAPORE

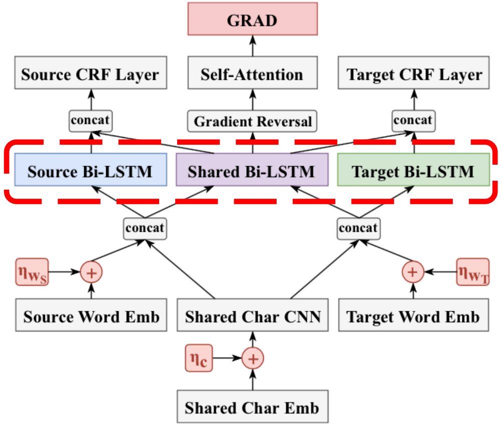
Representation Difference

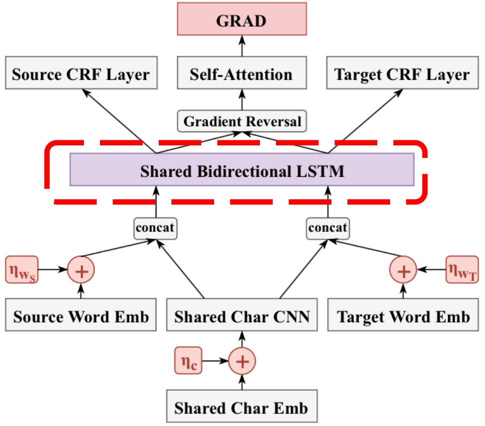
Partially Share (DATNet-P) and Fully Share (DATNet-F)

DATNet-P decomposes the BiLSTM units into the shared component and the private one.

In **DATNet-F**, the BiLSTM units are fully shared by both resources while word embeddings for different resources are disparate.

In the experiment, we will investigate the performance of two different shared representation architectures on different tasks and give their corresponding recommendation.





Agency for Science, Technology and Research
SINGAPORE

DATNet: Experiments

In this experiment, CoNLL-2003 English NER is source data, CoNLL-2002 and WNUT are target data.

Cross-language transfer: CoNLL-2003 → CoNLL-2002

Improvement 3%

Cross-domain transfer: CoNLL-2003 → WNUT

Improvement 6%

1. **DATNet-P** model advocates Cross-language.
2. **DATNet-F** model advocates **Cross-domain.**

Comparison with State-of-the-art Results in CoNLL and WNUT datasets (F1-score).

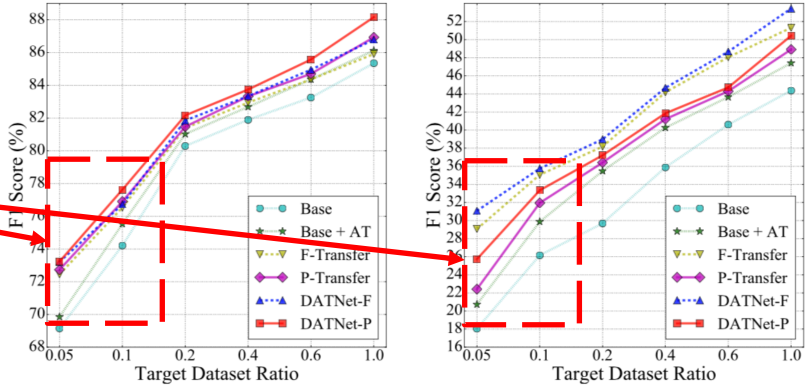
Mode	Methods		Additional Features			CoNLL Datasets		WNUT Datasets	
			POS	Gazetteers	Orthographic	Spanish	Dutch	WNUT-2016	WNUT-2017
Mono-language /domain	Gillick *et al.* [74]		×	×	×	82.59	82.84	-	-
	Lample *et al.* [4]		×	√	×	85.75	81.74	41.77*	34.53*
	Partalas *et al.* [67]		√	√	√	-	-	46.16	-
	Limsopatham *et al.* [68]		×	×	√	-	-	52.41	-
	Lin *et al.* [75]		√	√	×	-	-	-	40.42
	Our Base Model	Best	×	×	×	85.53	85.55	44.96	35.20
		Mean & Std				85.35±0.15	85.24±0.21	44.37±0.31	34.67±0.34
Cross-language /domain	Yang *et al.* [13]		×	√	×	85.77	85.19	47.19*	40.83*
	Ying *et al.* [35]		×	√	×	85.88	86.55	46.53*	40.79*
	Feng *et al.* [21]		√	×	×	86.42	**88.39**	-	-
	Von *et al.* [76]		×	√	×	-	-	-	40.78
	Aguilar *et al.* [33]		√	×	√	-	-	-	41.86
	DATNet-P	Best	×	×	×	**88.16**	88.32	50.85	41.12
		Mean & Std				87.89±0.18	88.09±0.13	50.41±0.32	40.52±0.38
	DATNet-F	Best	×	×	×	87.04	87.77	**53.43**	**42.83**
		Mean & Std				86.79±0.20	87.52±0.19	53.03±0.24	42.32±0.32

The scores with "*" denote produced results by the corresponding official tools/codes.

Agency for Science, Technology and Research
SINGAPORE

Transfer Learning Performance

- The transfer learning component in the DATNet consistently improves over the results of the base model and the improvement margin is more distinct when the **target data ratio is lower**.



(a) CoNLL-2002 Spanish NER (b) WNUT-2016 Twitter NER

Experiments on Extremely Low Resource (F1-score).

Tasks	CoNLL-2002 Spanish NER					
# Target sentences	10	50	100	200	500	1000
Base	21.53	42.18	48.35	63.66	68.83	76.69
+ AT	19.23	41.01	50.46	64.83	70.85	77.91
+ P-Transfer	29.78	61.09	64.78	66.54	72.94	78.49
+ F-Transfer	39.72	63.00	63.36	66.39	72.88	78.04
DATNet-P	39.52	62.57	64.05	**68.95**	**75.19**	**79.46**
DATNet-F	**44.52**	**63.89**	**66.67**	68.35	74.24	78.56
Tasks	WNUT-2016 Twitter NER					
# Target sentences	10	50	100	200	500	1000
Base	3.80	14.07	17.99	26.20	31.78	36.99
+ AT	4.34	16.87	18.43	26.32	35.68	41.69
+ P-Transfer	7.71	16.17	20.43	29.20	34.90	41.20
+ F-Transfer	15.26	20.04	26.60	32.22	38.35	44.81
DATNet-P	9.94	17.09	25.39	30.71	36.05	42.30
DATNet-F	**17.14**	**22.59**	**28.41**	**32.48**	**39.20**	**45.25**

- DATNet-F outperforms DATNet-P on cross-language transfer when the target resource is extremely low, however, this results are reversed when the target dataset size is large enough (i.e., more than 100 sentences);
- DATNet-F is generally superior to DATNet-P on cross-domain transfer.

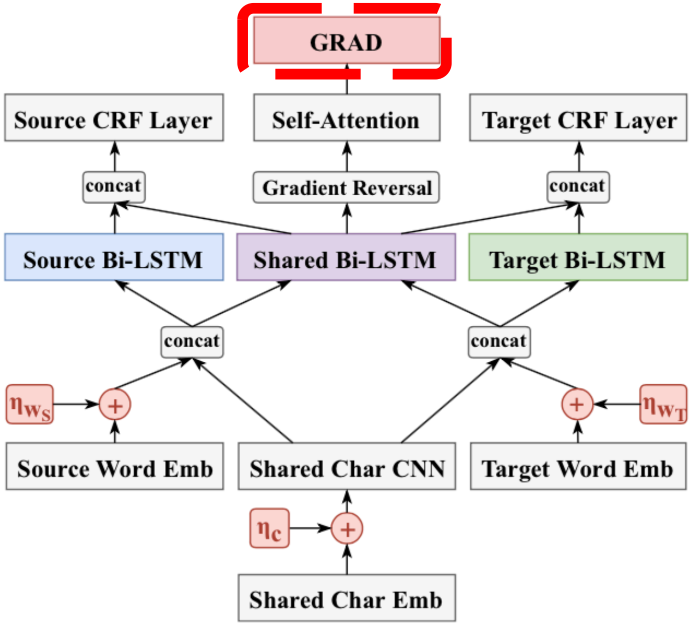
Resource Data Imbalance

Generalized Resource-Adversarial Discriminator (GRAD)

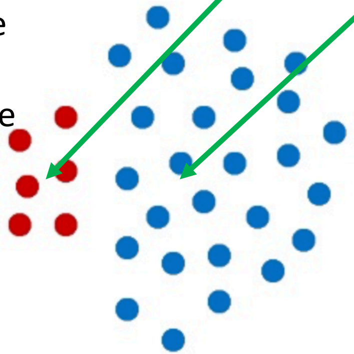
GRAD takes self-attention output and computes the resource label. Its loss is defined as

$$
\begin{aligned}
\ell_{GRAD} = -\sum_i \{ & \mathbf{I}_{i \in \mathcal{D}_S} \alpha (1 - r_i)^{\gamma} \log r_i \\
& + \mathbf{I}_{i \in \mathcal{D}_T} (1 - \alpha) r_i^{\gamma} \log(1 - r_i) \}
\end{aligned}
$$

α is a weighting factor to balance the loss contribution from high and low resource.



Resource Data Imbalance



α	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8
Ratio	CoNLL-2002 Spanish NER														
$\rho = 0.1$	78.37	78.63	**78.70**	78.32	77.96	77.92	77.88	77.78	77.85	77.90	77.65	77.57	77.38	77.49	77.29
$\rho = 0.2$	80.99	81.71	**82.18**	81.57	81.53	81.55	81.44	81.25	81.32	81.16	81.02	81.16	80.63	80.79	80.54
$\rho = 0.4$	83.76	83.73	84.18	**84.48**	84.26	84.12	83.54	83.40	83.52	84.18	83.42	83.47	83.28	83.33	83.19
$\rho = 0.6$	85.18	85.24	85.85	85.68	85.84	**86.10**	85.71	85.74	85.42	85.60	85.20	85.40	85.26	85.24	84.98

Table 5: Analysis of Discriminator Weight α in GRAD with Varying Data Ratio ρ (F1-score).

[1]Focal Loss for Dense Object Detection, ICCV 2017

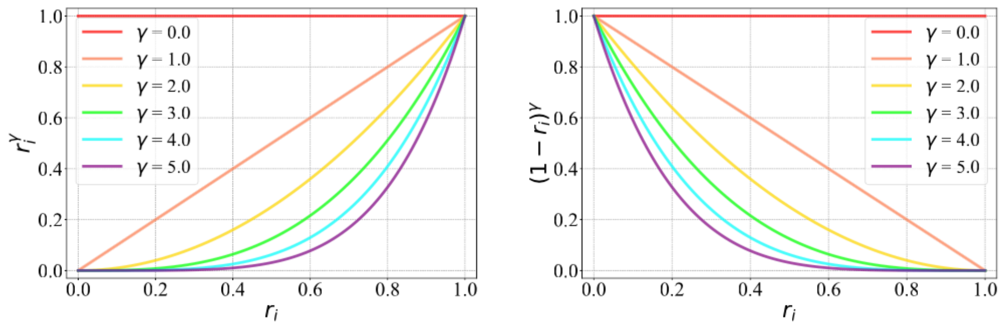
Resource Data Imbalance

Generalized Resource-Adversarial Discriminator (GRAD)

$$\ell_{GRAD} = -\sum_i \{\mathbf{I}_{i\in\mathcal{D}_S}\alpha\boxed{(1-r_i)^\gamma}\log r_i + \mathbf{I}_{i\in\mathcal{D}_T}(1-\alpha)\boxed{r_i^\gamma}\log(1-r_i)\}$$

$\boxed{(\mathbf{1}-\boldsymbol{r_i})^{\boldsymbol{\gamma}}\ \textbf{(or}\ \boldsymbol{r_i^\gamma}\textbf{)}}$ **controls the loss contribution from individual samples** by measuring the discrepancy between prediction and true label (easy samples have smaller contribution).

- For the sample from the high resource D_s, its corresponding loss term is $I_{i\in D_s}\alpha(1-r_i)^\gamma \log r_i$, where the controlling factor $(\mathbf{1}-\boldsymbol{r_i})^{\boldsymbol{\gamma}}$ **is inverse proportion to** $\boldsymbol{r_i}$. In other words, $r_i \to 1$, this well-classified sample is down-weighted due to $(1-r_i)^\gamma$ goes to 0. As γ increases, the approaching speed increases. In this case, **for sample from high resource data, a large $\boldsymbol{\gamma}$ is preferred**.
- On the contrary, **for the sample from low resource data, a small $\boldsymbol{\gamma}$ is preferred.**



(a) r_i vs r_i^γ (b) r_i vs $(1-r_i)^\gamma$

Agency for Science, Technology and Research
SINGAPORE

Ablation Study of GRAD

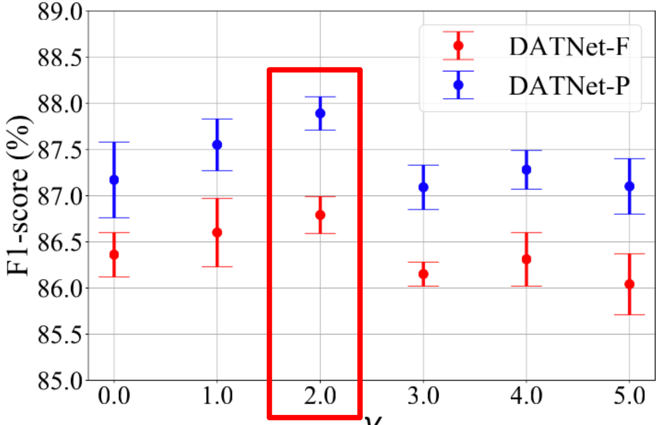
This experiment reports the quantitative performance comparison between models with different components.

1. **GRAD shows the stable superiority over the normal AD regardless of other components.**

Quantitative Performance Comparison between Models with Different Components.

Model	F1-score	Model	F1-score
CoNLL-2002 Spanish NER			
Base	85.35	+AT	86.12
+P-T (no AD)	86.15	+AT +P-T (no AD)	86.90
+F-T (no AD)	85.46	+AT +F-T (no AD)	86.17
+P-T (AD)	86.32	+AT +P-T (AD)	87.19
+F-T (AD)	85.58	+AT +F-T (AD)	86.38
+P-T (GRAD)	86.93	**DATNet-P**	**88.16**
+F-T (GRAD)	85.91	**DATNet-F**	87.04
WNUT-2016 Twitter NER			
Base	44.37	+AT	47.41
+P-T (no AD)	47.66	+AT +P-T (no AD)	48.44
+F-T (no AD)	49.79	+AT +F-T (no AD)	50.93
+P-T (AD)	48.14	+AT +P-T (AD)	49.41
+F-T (AD)	50.48	+AT +F-T (AD)	51.84
+P-T (GRAD)	48.91	**DATNet-P**	50.85
+F-T (GRAD)	51.31	**DATNet-F**	**53.43**

* AT: Adversarial Training; P-T: P-Transfer; F-T: F-Transfer; AD: Adversarial Discriminator; GRAD: Generalized Resource-Adversarial Discriminator.



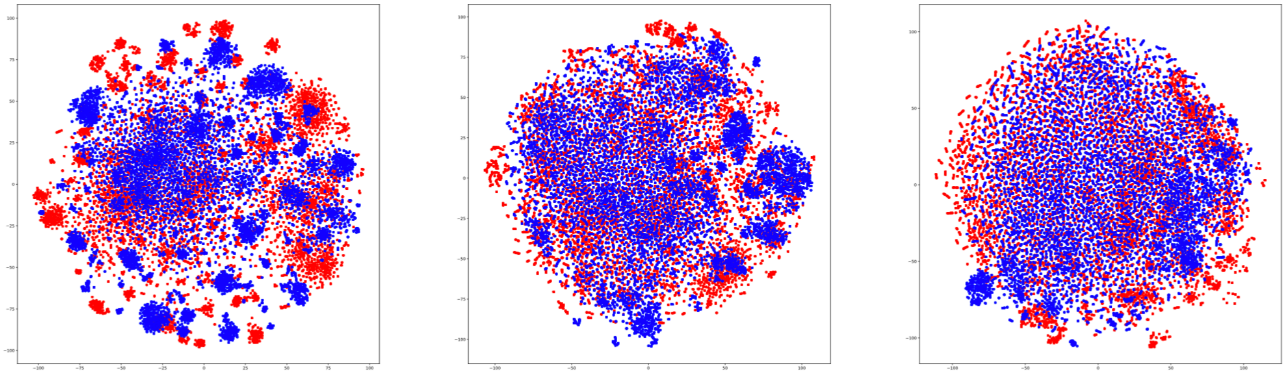
. Analysis of γ in GRAD on CoNLL-2002 Spanish NER.

The recommendation of $\gamma = 2$ for GRAD in practical use.

DATNet: Feature Visualization

The visualization of extracted features from shared bidirectional-LSTM layer. The left, middle, and right figures show the results when no Adversarial Discriminator (AD), AD, and GRAD is performed, respectively. Red points correspond to the source CoNLL-2003 English examples, and blue points correspond to the target CoNLL-2002 Spanish examples.

GRAD in DATNet makes the distribution of extracted features from the source and target datasets much more similar by considering the data imbalance, which indicates that ***the outputs of BiLSTM are resource-invariant***.



(a) no AD (b) AD (c) GRAD

Agency for Science, Technology and Research
SINGAPORE

DATNet: Adversarial Training (AT)

Adversarial samples are widely incorporated into training to **improve the generalization and robustness of the model**, which is called adversarial training.

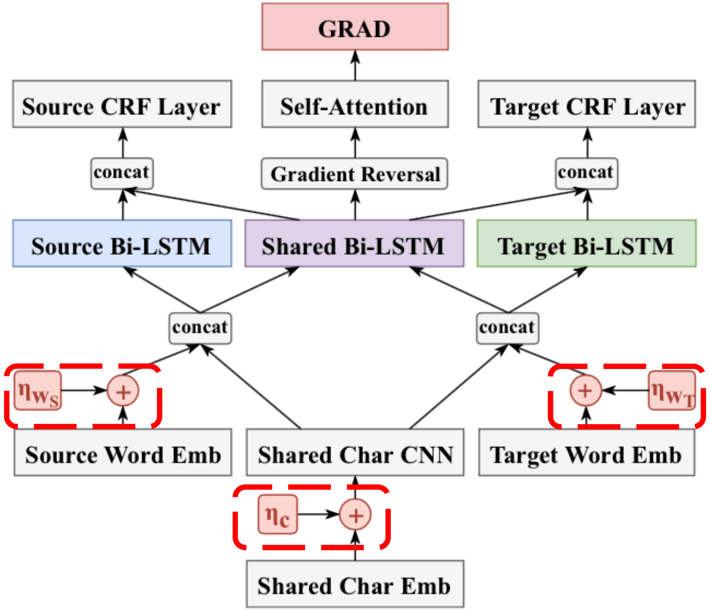
It emerges as a powerful regularization tool to **stabilize training and enable the model to escape from the local minimum**.

an adversarial sample is built by adding the original sample with a perturbation bounded by a small norm ϵ to maximize the loss function as

$$\eta_{\mathbf{x}} = \arg \max_{\eta:\|\eta\|_2 \leq \epsilon} \ell(\Theta; \mathbf{x} + \eta)$$

where Θ is the current model parameters set. η is estimated by

$$\eta_{\mathbf{x}} = \epsilon \frac{\mathbf{g}}{\|\mathbf{g}\|_2}, \quad \text{where } \mathbf{g} = \nabla \ell(\Theta; \mathbf{x})$$



Agency for Science, Technology and Research SINGAPORE

DATNet: Experiments Ablation Study

The aforementioned results show AT helps to enhance the overall performance by adding perturbations into inputs with the limit of $\epsilon = 5$.

This experiment indicates that less **training data require a larger ϵ to prevent over-fitting**, which further validates the necessity of AT in the case of low resource data.

Analysis of Maximum Perturbation $\epsilon_{\mathbf{w}_T}$ in AT with Varying Data Ratio ρ (F1-score).

$\epsilon_{\mathbf{w}_T}$	1.0	3.0	5.0	7.0	9.0
Ratio	CoNLL-2002 Spanish NER				
$\rho = 0.1$	75.90	76.23	77.38	77.77	**78.13**
$\rho = 0.2$	81.54	81.65	81.32	**81.81**	81.68
$\rho = 0.4$	83.62	83.83	83.43	**83.99**	83.40
$\rho = 0.6$	84.44	84.47	**84.72**	84.04	84.05

Agency for Science, Technology and Research
SINGAPORE

Agency for Science, Technology and Research
SINGAPORE

CREATING GROWTH, ENHANCING LIVES

Thank you