肺結節偵測基於多重注意力機制與多尺度特徵融合之殘差架構U-Net

Lung Nodule Detection Based on The Residual U-Net with Multi-attention and Multi-scale Feature Fusion

研究生：NM6104088 賴亭諭
指導教授：徐禕佑 助理教授

大綱

1. 介紹
1.1 研究動機
1.2 研究目標
1.3 研究貢獻
2. 訓練環境和資料集
2.1 訓練設備
2.2 資料集
2.3 資料集前處理
3. 相關研究
3.1 U-Net[7]
3.2 Residual unit [10]
3.3 scSE [14]
3.4 Vision transformer (ViT) [11]
3.5 多尺度特徵融合[12]
4. 模型結構
4.1 殘差單元 — 模型1
4.2 scSE — 模型1
4.3 多尺度特徵融合— 模型1
4.4 Vision transformer (ViT) — 模型1
4.5 scSE — 模型2
4.6 損失函數
5. 實驗結果
5.1 針對注意力機制之消融研究
5.2 針對ViT之研究
5.3 針對影像剪切大小之研究
5.4 與其他方法之比較
5.5 本論文二種模型之比較
6. 結論
7. 參考資料
附錄

1.介紹

研究動機

111年死因前三名

1.惡性腫瘤(癌症)

2.心臟疾病

3.嚴重特殊傳染性肺炎(COVID-19)

111年癌症死亡率前三名

1.氣管、支氣管和肺癌

2.肝和肝內膽管癌

3.結腸、直腸和肛門癌

肺結節

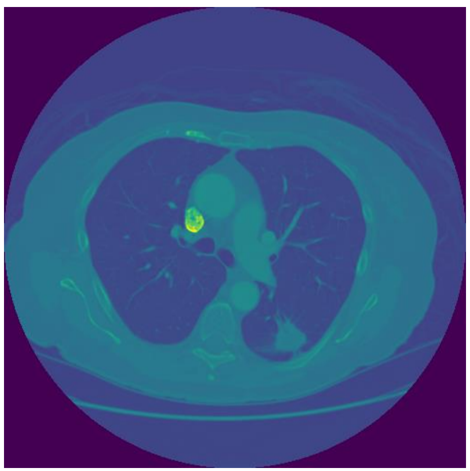
- 肺癌初期症狀
- 不易判讀
- 成因與發展廣泛，不一定會發展成肺癌
- 50歲以上的人，有三分之二都有肺結節

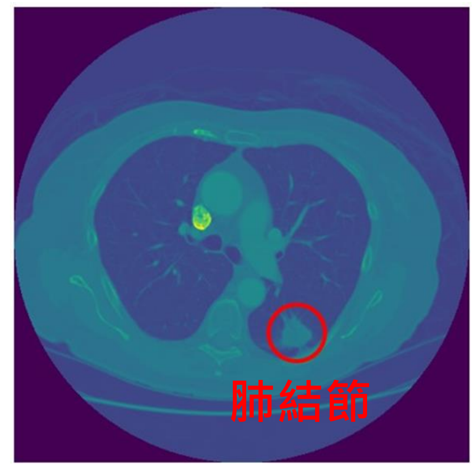
111年國人死因統計結果：https://www.mohw.gov.tw/cp-16-74869-1.html

研究目標

創建模型

- 自動擷取特徵並學習圖樣
- 以LIDC-IDRI為資料集來源
- 用scSE[14]注意力機制、ViT [11]與多尺度特徵融合強化U-Net





[11]An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
[14] Concurrent spatial and channel 'squeeze and excitation' in fully convolutional networks

研究貢獻

1. 相較於U-Net、Attention Unet、TransUNet，本研究模型成功提高IoU達0.82，並且降低39.61%假陰性樣本，提高肺結節存在與否之判讀準確度至94.63%
2. 加快醫師診斷速度
3. 以消融實驗實證，在原始論文[14]表現較佳的模式並不適用於所有狀況，模型須因應應用場景的不同做出相對應的調整。因此，在本研究中，將注意力機制[14]擺放位置進行調整。
4. 提出2種模型供使用者選擇

[14] Concurrent spatial and channel 'squeeze and excitation' in fully convolutional networks

2. 訓練環境和資料集

訓練設備

- 廣達qpm運算平台（Nvidia A100-MIG-3g.40gb）

資料集

- 肺影像資料庫聯盟和影像資料庫資源倡議（The lung image database consortium and image database resource initiative, LIDC-IDRI）
- 美國國立癌症研究所蒐集7個不同醫療中心1018名病人的低劑量CT
- LIDC-IDRI資料集的內容主要包括以下內容：
  1. CT影像
  2. 醫學專家標註
  3. 臨床和影像資訊
  4. 評估和評分

LIDC-IDRI資料集載點：https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=1966254

資料集前處理

- 步驟 1. 轉換影像單位

$$HU = pixel\ value \times rescale\ slope + rescale\ intercept$$

```
Dataset.file_meta -------------------------------
(0002, 0000) File Meta Information Group Length  UL: 206
(0002, 0001) File Meta Information Version       OB: b'\x00\x01'
(0002, 0002) Media Storage SOP Class UID         UI: CT Image Storage
(0002, 0003) Media Storage SOP Instance UID      UI: 1.3.6.1.4.1.14519.5.2.1.6279.6001.806997748376560324284235498426
(0002, 0010) Transfer Syntax UID                 UI: Explicit VR Little Endian
(0002, 0012) Implementation Class UID            UI: 1.3.6.1.4.1.22213.1.143
(0002, 0013) Implementation Version Name         SH: '0.5'
(0002, 0016) Source Application Entity Title     AE: 'POSDA'
-------------------------------------------------
(0008, 0005) Specific Character Set              CS: 'ISO_IR 100'
(0008, 0008) Image Type                          CS: ['ORIGINAL', 'PRIMARY', 'AXIAL']
(0008, 0016) SOP Class UID                       UI: CT Image Storage
(0008, 0018) SOP Instance UID                    UI: 1.3.6.1.4.1.14519.5.2.1.6279.6001.806997748376560324284235498426
(0008, 0020) Study Date                          DA: '20000101'
(0008, 0021) Series Date                         DA: '20000101'
(0008, 0022) Acquisition Date                    DA: '20000101'
(0008, 0023) Content Date                        DA: '20000101'
(0008, 0024) Overlay Date                        DA: '20000101'
(0008, 0025) Curve Date                          DA: '20000101'
(0008, 002a) Acquisition DateTime                DT: '20000101'
(0008, 0030) Study Time                          TM: ''
(0008, 0032) Acquisition Time                    TM: ''
(0008, 0033) Content Time                        TM: ''
(0008, 0050) Accession Number                    SH: ''
...
(0040, a124) UID                                 UI: 1.3.6.1.4.1.14519.5.2.1.6279.6001.178367864784638350333907738265
(0070, 0084) Content Creator's Name              PN: ''
(0088, 0140) Storage Media File-set UID          UI: 1.3.6.1.4.1.14519.5.2.1.6279.6001.228028567290201938725448638500
(7fe0, 0010) Pixel Data                          OW: Array of 524288 elements
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

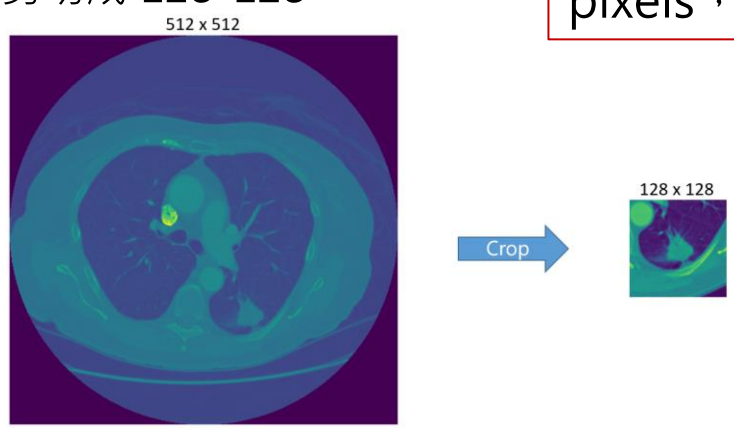
```
(0028, 1052) Rescale Intercept                   DS: '-1024.0'
(0028, 1053) Rescale Slope                       DS: '1.0'
```

轉換HU值優點：
1. 消除影像間的差異
2. 提供組織密度資訊
3. 減少雜訊和增強對比

- 步驟 2. 剪切影像大小

將原始影像（512*512）剪切成 128*128

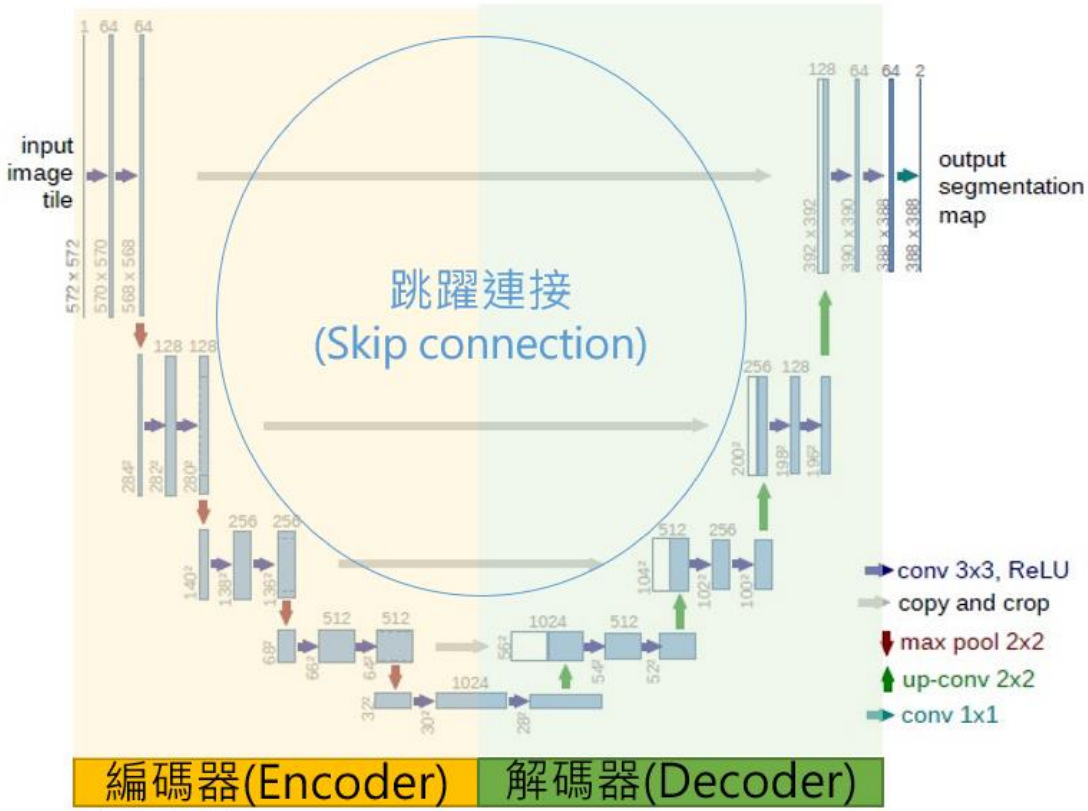
資料集中最長的肺結節為49 pixels，並無超出影像的問題。



3. 相關研究

U-Net

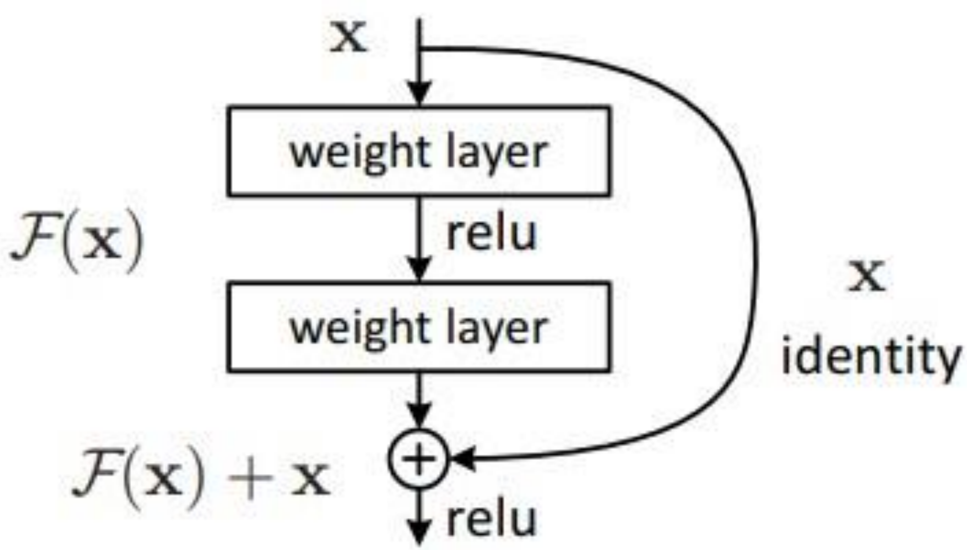
- 「編碼器—解碼器」結構
- 跳躍連接

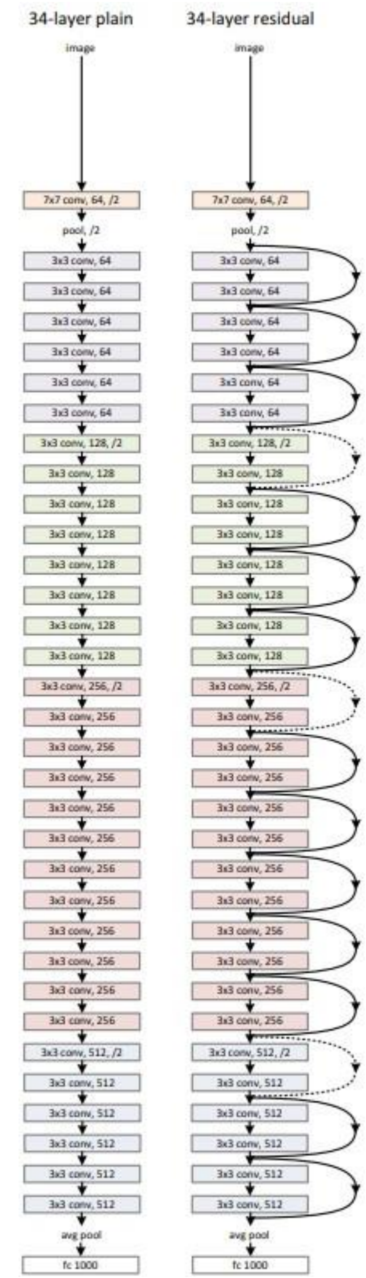


[7]U-net: Convolutional networks for biomedical image segmentation

Residual U-Net

- 解決深度神經網路中的梯度消失問題
- 簡化模型訓練
- 有助於訊息傳播，且不會降低性能





[10]Deep Residual Learning for Image Recognition

scSE

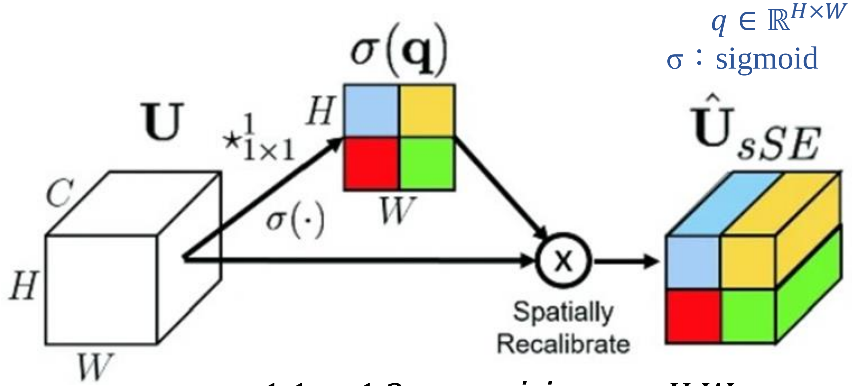
通道擠壓與空間激發

$u^{i,j} \in \mathbb{R}^{1\times1\times C}$

$W_{sq} \in \mathbb{R}^{1\times1\times C\times1}$

$q \in \mathbb{R}^{H\times W}$

σ：sigmoid

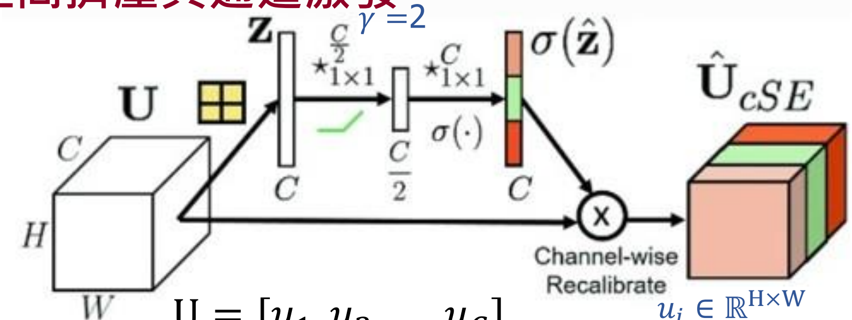


$$U = [u^{1,1}, u^{1,2}, \dots, u^{i,j}, \dots, u^{H,W}]$$

$$q = W_{sq} * U$$

$$\widehat{U}_{sSE} = [\sigma(q_{1,1})u^{1,1}, \sigma(q_{1,2})u^{1,2}, \dots, \sigma(q_{i,j})u^{i,j}, \dots, \sigma(q_{H,W})u^{H,W}]$$

空間擠壓與通道激發



$$U = [u_1, u_2, \dots, u_C]$$

Channel Descriptor

$$z_k = \frac{1}{H\times W}\sum_{i}^{H}\sum_{j}^{W} u_k(i,j)$$

$$\hat{z} = W_1(\delta(W_2 z_k))$$

$$\widehat{U}_{cSE} = [\sigma(\widehat{z_1})u_1, \sigma(\widehat{z_2})u_2, \dots, \sigma(\widehat{z_C})u_C]$$

$u_i \in \mathbb{R}^{H\times W}$

$z \in \mathbb{R}^{1\times1\times C}$

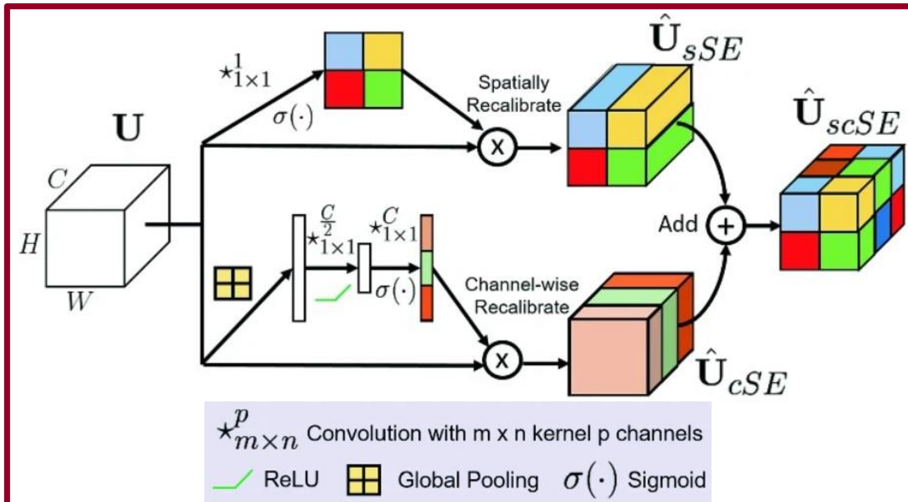
$W_1 \in \mathbb{R}^{C\times\frac{C}{\gamma}}$

$W_2 \in \mathbb{R}^{\frac{C}{\gamma}\times C}$

γ：Reduction Ratio

σ：sigmoid

δ：ReLU

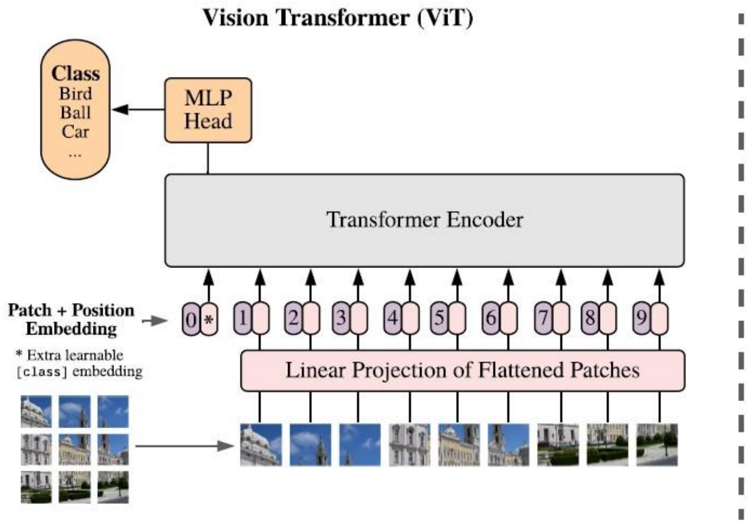


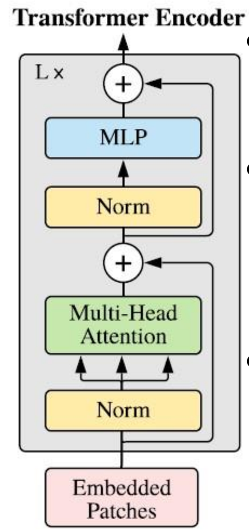
$$\widehat{U}_{scSE} = \widehat{U}_{sSE} + \widehat{U}_{cSE}$$

- scSE 優點：
  (1) 自適應調整權重
  (2) 減少過擬合
  (3) 可以輕易嵌入原有的CNN架構
  (4) 提高模型性能

[14] Concurrent spatial and channel 'squeeze and excitation' in fully convolutional networks

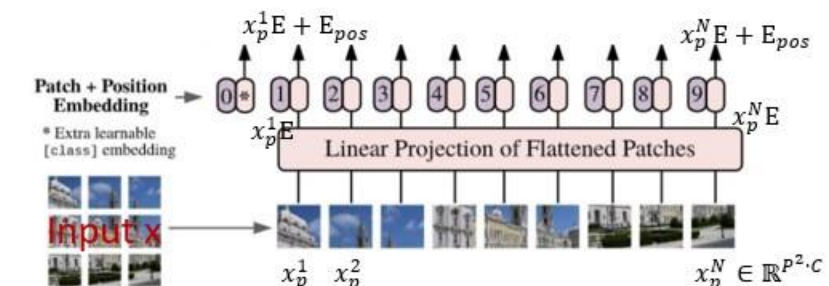
Vision transformer (ViT)





- ViT的基本組成單元是多層自注意力層
- ViT的訓練過程包括兩個主要步驟：Patch Embedding和Transformer Training
- ViT在處理大規模和複雜影像有出色表現

Patch Embedding



$X \in \mathbb{R}^{H \times W \times C}$

$N = \dfrac{H \times W}{P^2}$

$$X = [x_P^1, x_P^2, \dots, x_P^N]$$

$E \in \mathbb{R}^{(P^2 \times C) \times D}$

Learnable Embedding

Position Embedding

$$z_0 = [x_{Class}; x_P^1E; x_P^2E; \dots; x_P^NE] + E_{pos}$$

Transformer Training

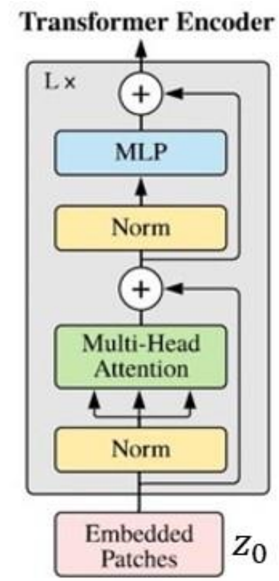
Multiheaded Self-attention

$$z'_\ell = MSA(LN(z_{\ell-1})) + z_{\ell-1}$$

$$z_\ell = MLP(LN(z'_\ell)) + z'_\ell$$

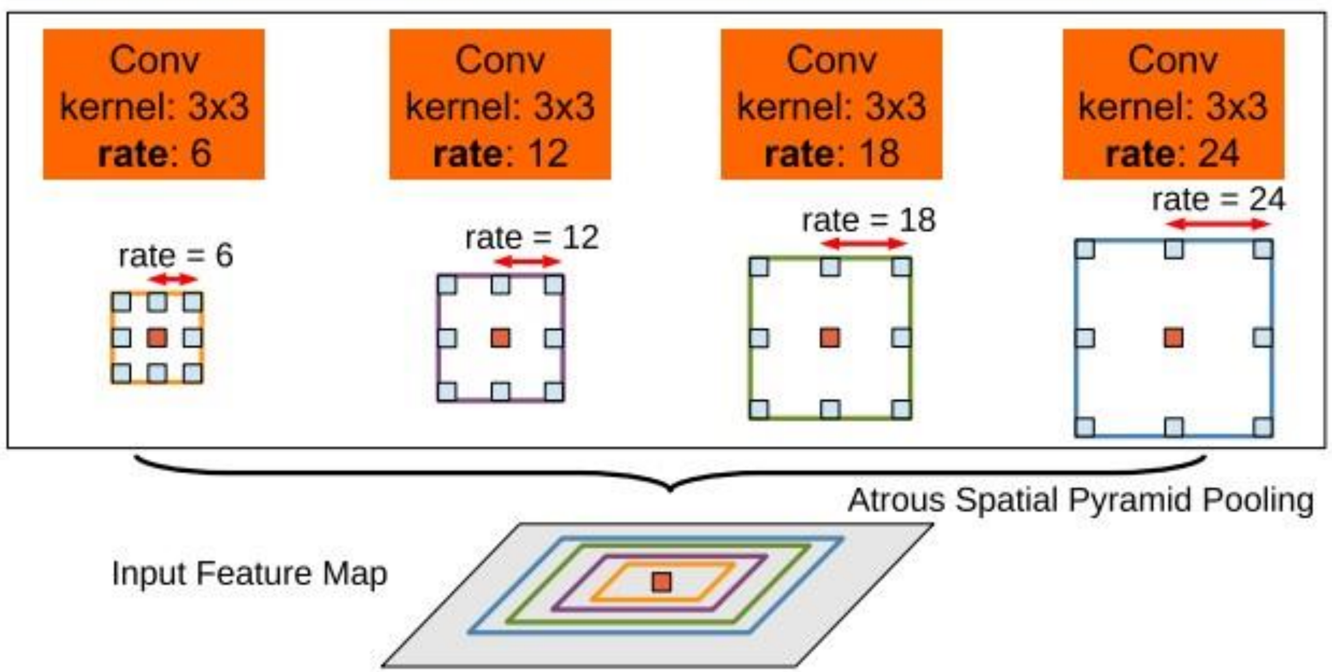
Multi-Layer Perceptron

Layer Norm



多尺度特徵融合

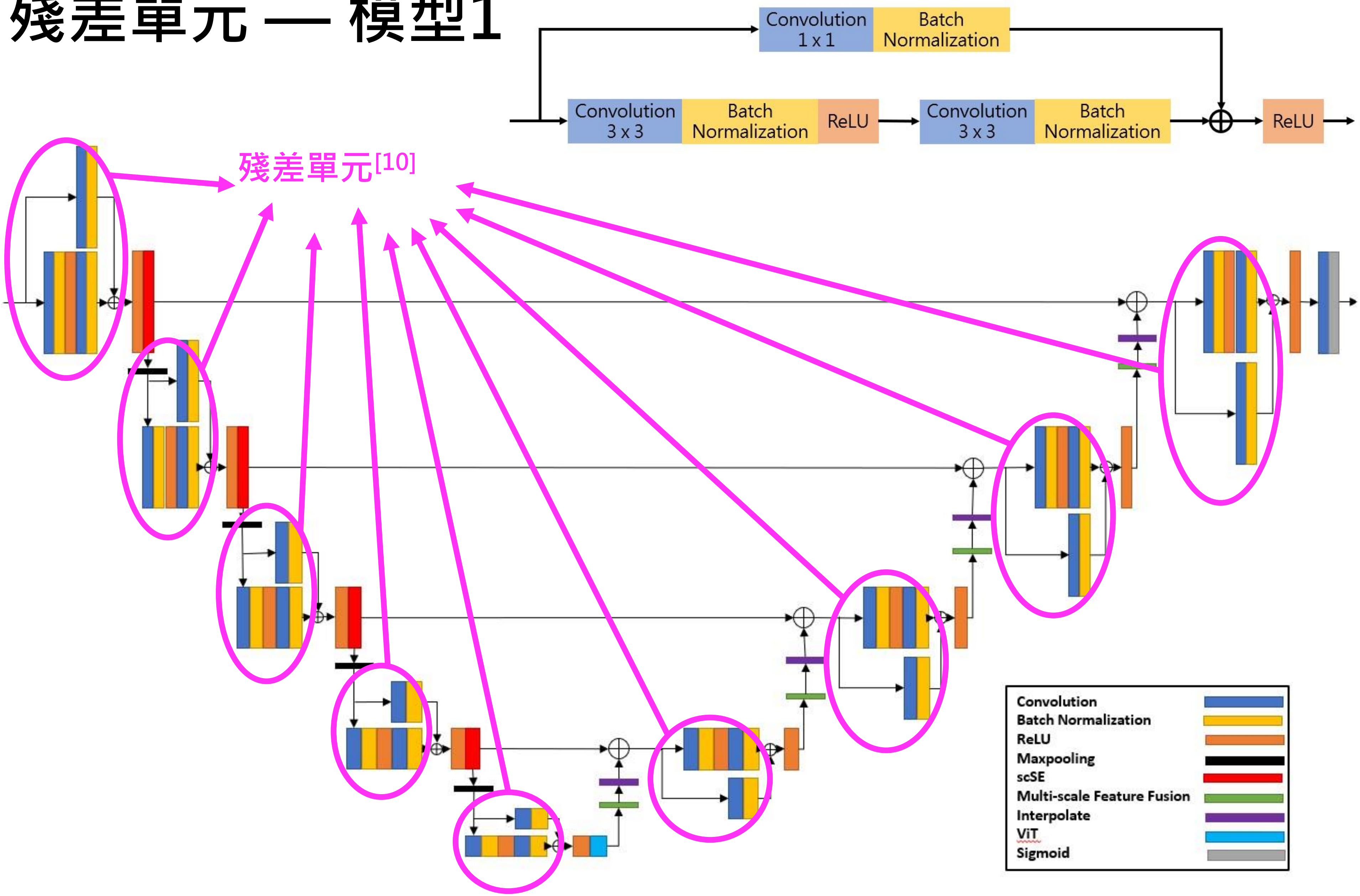
- 多個不同擴張率的空洞卷積組成
- 不增加額外的計算量和記憶體消耗的情況下，有效地擴展模型對上下文的理解能力



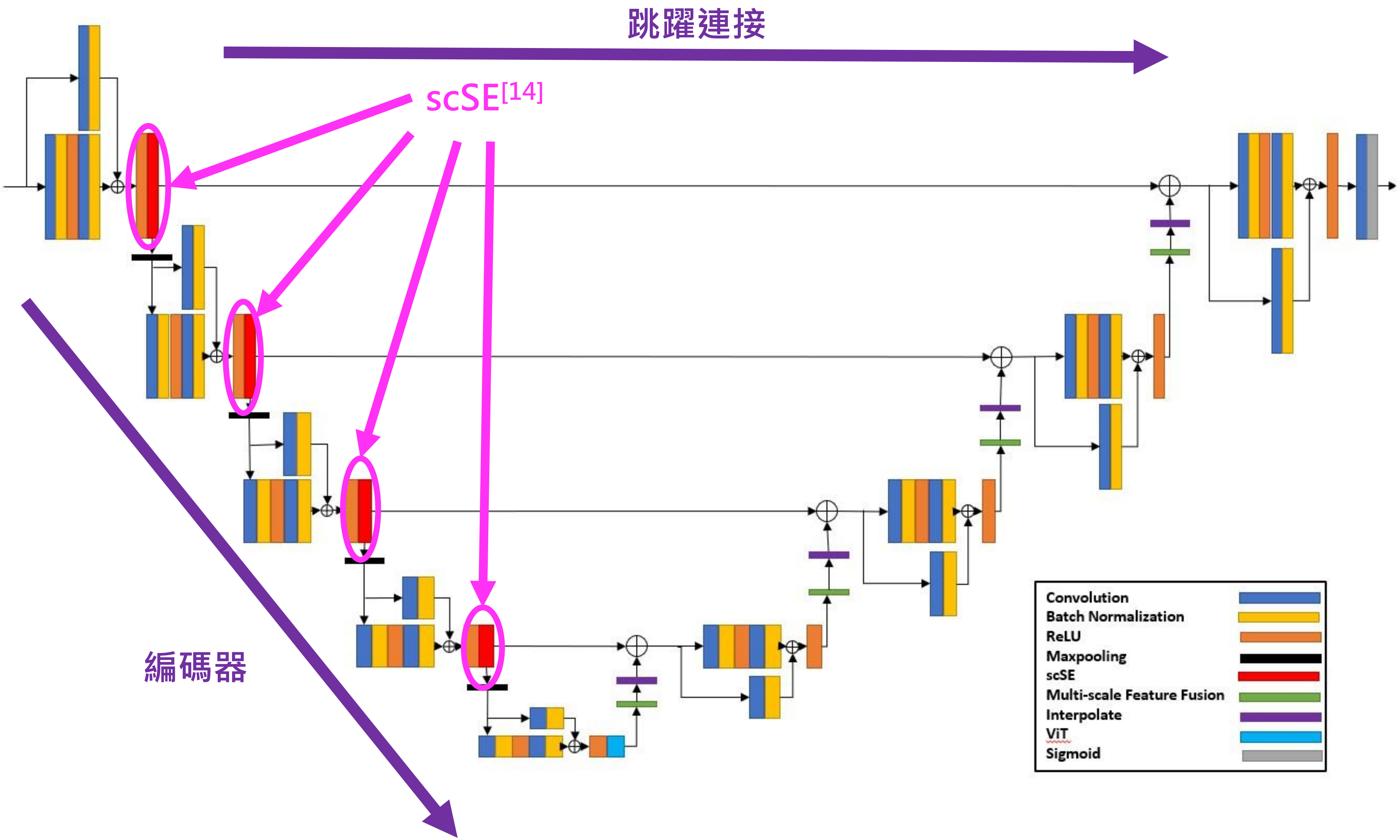
[12]DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs

4. 模型結構

殘差單元 — 模型1

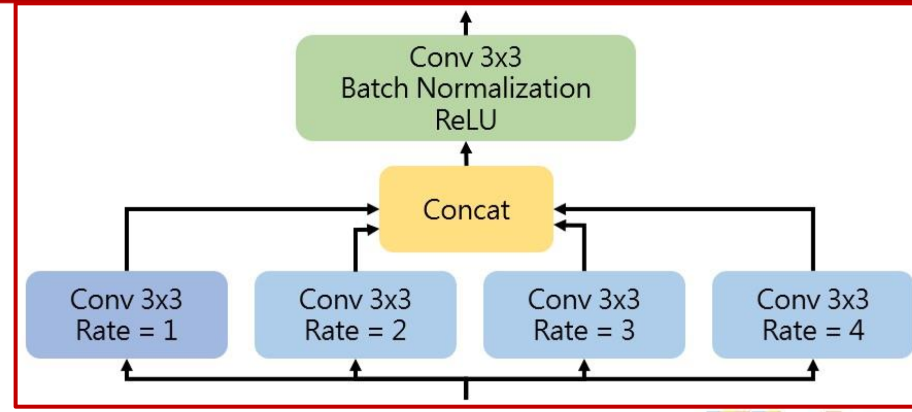

Convolution
1 x 1
Batch
Normalization
Convolution
3 x 3
Batch
Normalization
ReLU
Convolution
3 x 3
Batch
Normalization
ReLU
殘差單元[10]
Convolution
Batch Normalization
ReLU
Maxpooling
scSE
Multi-scale Feature Fusion
Interpolate
ViT
Sigmoid

scSE — 模型1

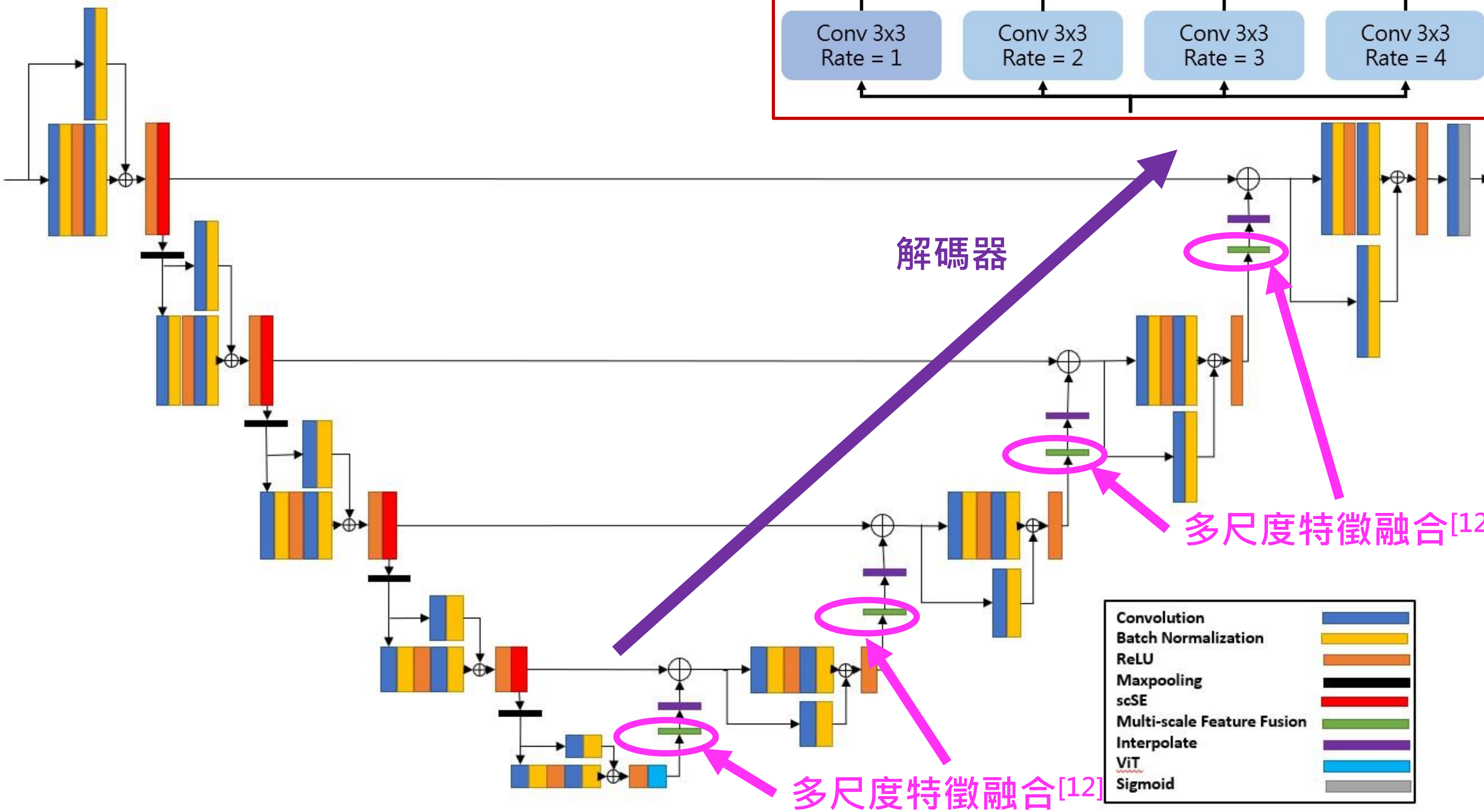


跳躍連接
scSE[14]
編碼器
Convolution
Batch Normalization
ReLU
Maxpooling
scSE
Multi-scale Feature Fusion
Interpolate
ViT
Sigmoid

多尺度特徵融合 — 模型1

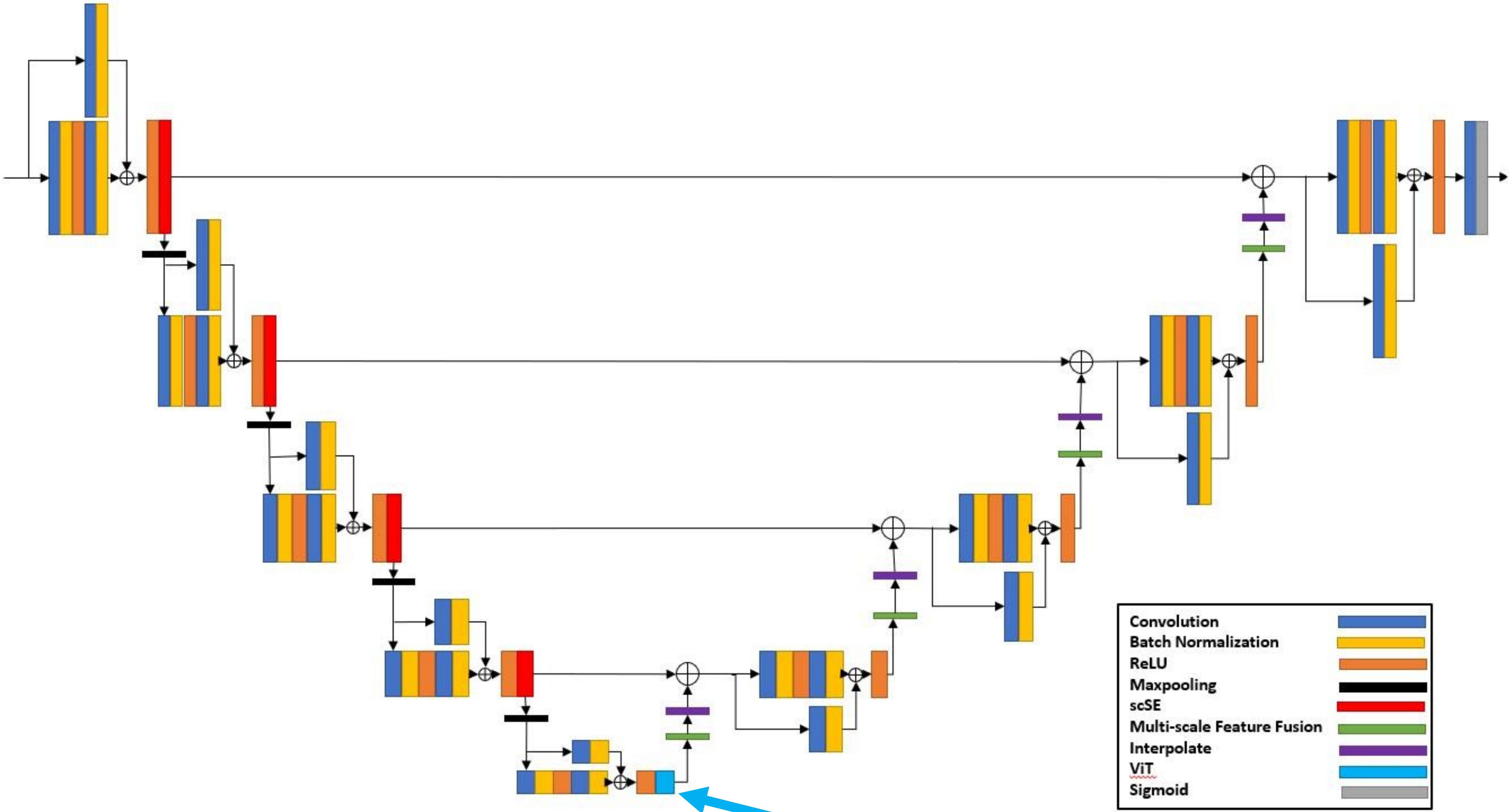


Conv 3x3
Batch Normalization
ReLU
Concat
Conv 3x3
Rate = 1
Conv 3x3
Rate = 2
Conv 3x3
Rate = 3
Conv 3x3
Rate = 4



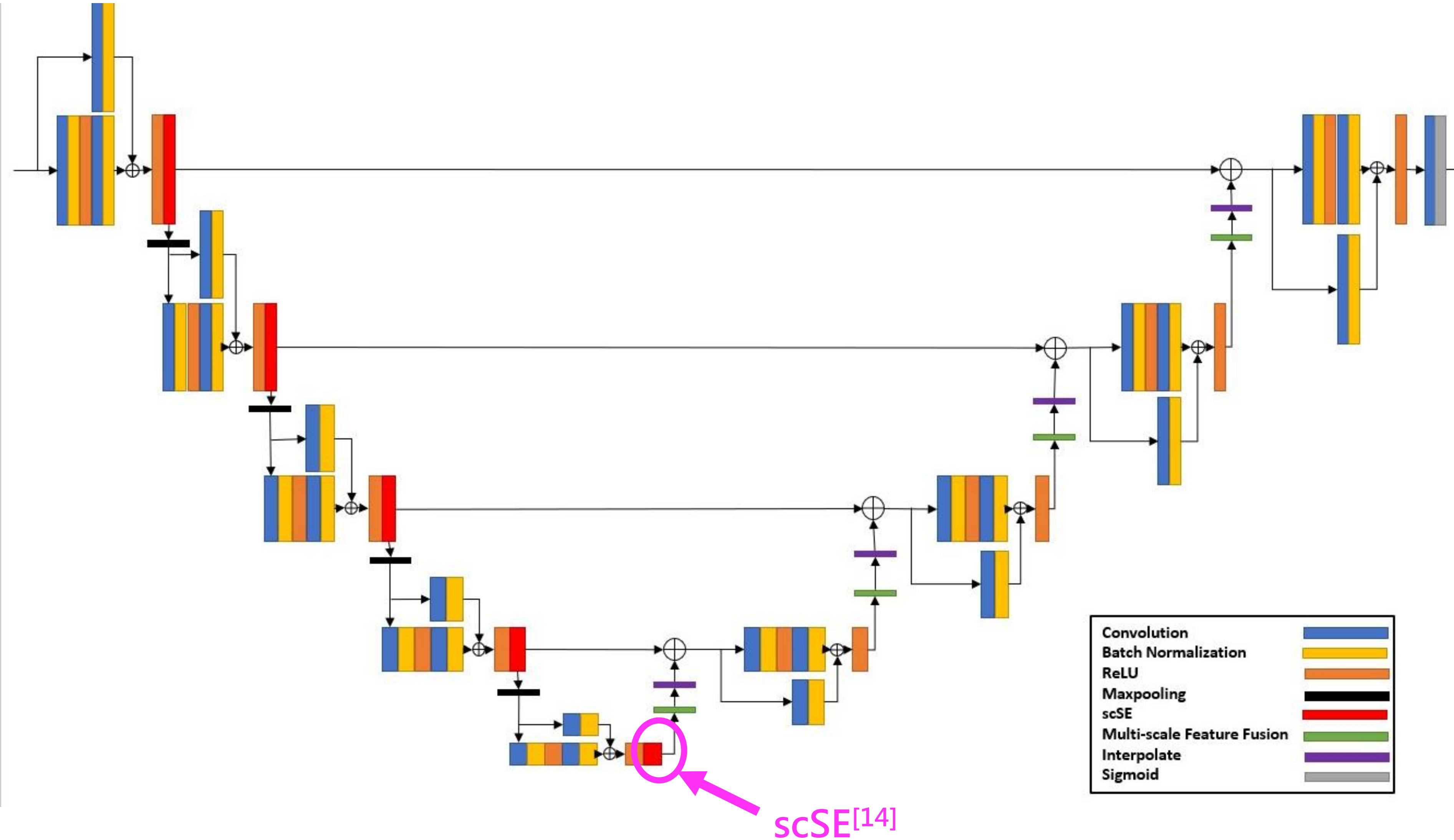
解碼器
多尺度特徵融合[12]
多尺度特徵融合[12]
Convolution
Batch Normalization
ReLU
Maxpooling
scSE
Multi-scale Feature Fusion
Interpolate
ViT
Sigmoid

ViT — 模型1



Vision transformer (ViT) [11]

scSE — 模型2



Convolution
Batch Normalization
ReLU
Maxpooling
scSE
Multi-scale Feature Fusion
Interpolate
Sigmoid
scSE[14]

損失函數

- Dice loss對於正負樣本嚴重不平衡有很好的表現

$$Dice\ loss = 1 - \frac{2|X \cap Y|}{|X| + |Y|}$$

Dice係數

- Binary cross-entropy loss（BCE loss）根據實際標籤和預測結果之間的差異來衡量模型的預測性能

$$BCE\ loss = -\frac{1}{N}\sum_{n=1}^{N} w_n[y_n \log x_n + (1 - y_n)\log(1 - x_n)]$$

- 結合BCE loss和Dice loss是為了要同時考慮預測的精確度和分割相似度

$$\boldsymbol{Loss = 0.5 \times BCE\ loss + Dice\ loss}$$

5. 實驗結果

針對注意力機制之消融研究

- **實驗設定**：4094張測試影像並且模型已加入殘差單元和多尺度特徵融合
- **實驗目標**：本實驗以Concurrent spatial and channel 'squeeze and excitation' in fully convolutional networks[14]論文為基礎，找出scSE最佳擺放位置。

scSE Position	IoU	Confusion matrix			
		TP	TN	FP	FN
None	0.740	2452	1373	4	265
Only Encoder	**0.819**	2447	1373	4	270
Only Decoder	0.812	2438	1369	8	279
Only Bottleneck	0.811	2432	1368	9	285
Encoder + Bottleneck	0.818	2481	1369	8	**236**
Decoder + Bottleneck	0.806	2403	1368	9	314
Encoder + Decoder	0.815	2474	1369	8	243
Encoder + Bottleneck + Decoder	0.816	2456	1370	7	261

針對ViT之研究

- **實驗動機**：scSE [14]在Only Encoder狀態下會有較高的IoU，而在Encoder + Bottleneck狀態下會有較好的Confusion matrix，假使在Bottlenck加入ViT [11]，會不會讓模型兼具高IoU和Confusion matrix
- **實驗設定**：4094張測試影像並且模型已加入殘差單元和多尺度特徵融合

scSE Position	ViT	IoU	Confusion Matrix				
			TP	TN	FP	FN	Acc.(%)
Only Encoder 模型1	No	0.819	2447	1373	4	270	93.31
Only Encoder	Bottleneck	**0.820**	**2510**	**1364**	**14**	**207**	**94.63**
Encoder + Bottleneck	No	0.818	2481	1369	8	236	94.04

模型2

針對影像剪切大小之研究

- **實驗動機**：部分論文有提到會先對影像做剪切才進行訓練，若影像剪裁越小是否能得到更好的結果？
- **實驗目標**：找出訓練影像時的最佳剪切尺寸
- **實驗設定**：4094張測試影像並且模型已加入殘差單元和多尺度特徵融合

Pixel size	Dice	IoU	Confusion Matrix				
			TP	TN	FP	FN	Accuracy
512×512	0.69	0.625	1385	1353	24	1332	0.669
256×256	0.872	0.810	2431	1368	9	286	0.928
128×128	**0.879**	**0.820**	**2510**	**1364**	**14**	**207**	**0.946**
64×64	0.875	0.812	2468	1366	11	249	0.936

與其他方法之比較

- **實驗目標**：確認本論文提出之二種模型有無比現存模型更好

Network	IoU	Confusion Matrix				
		TP	TN	FP	FN	Accuracy
U-Net	0.781	2462	1340	37	255	92.9%
Attention U-Net	0.814	2450	1368	9	267	93.3%
TransUnet	0.808	2428	1364	13	289	92.6%
scSE (Encoder) + ViT (Bottleneck)	**0.820**	**2510**	**1364**	**14**	**207**	**94.6%**
scSE (Encoder + Bottleneck)	0.818	2481	1369	8	236	94.0%

本論文二種模型之比較

- **實驗目標**：比較本論文提出之二種模型參數，分析所適用情境

Network	Interference Time	Parameter (個)	Params Size	IoU	Confusion matrix				
					TP	TN	FP	FN	Acc.
模型1	23.4 ms	115,596,353	400.97 MB	**0.820**	2510	1364	14	207	**94.6%**
模型2	**15.3 ms**	**71,538,433**	**272.9 MB**	0.818	2481	1369	8	236	94.0%

6. 結論

- 第一種同時使用scSE[14]和ViT作為注意力機制，將scSE [14]放置於Encoder各層，ViT [11]放置於Bottleneck，實驗結果得出判斷肺結節範圍位置的IoU達到0.82，且判斷是否存在肺結節的準確率為94.63%，模型推理時間需23.4ms，模型參數大小為400.97MB

- 第二種只使用scSE [14]作為注意力機制，將scSE [14]放置於Encoder各層和Bottleneck，判斷肺結節範圍位置的IoU為0.818，且判斷是否存在肺結節的準確率為94.0%，模型推理時間需15.3ms，模型參數大小為272.9MB。

- 簡而言之，若使用時需要較高IoU和較低假陰性時，則可以優先考慮使用scSE（Encoder）+ ViT（Bottleneck）架構。若有時間與記憶體限制，則可以考慮使用scSE（Encoder+bottleneck）架構。

7. 參考資料

[1] 維基百科. "癌症." https://zh.wikipedia.org/zh-tw/%E7%99%8C%E7%97%87 (accessed.

[2] 衛生福利部統計處. "110年國人死因統計結果." https://www.mohw.gov.tw/cp-16-70314-1.html (accessed.

[3] 中華民國內政部. "特定死因除外簡易生命表." https://www.moi.gov.tw/cl.aspx?n=2948 (accessed.

[4] A. C. o. Radiology. "Lung CT Screening Reporting & Data System (Lung-RADS®)." https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Lung-Rads (accessed.

[5] 維基百科. "亨氏單位." https://zh.wikipedia.org/wiki/%E4%BA%A8%E6%B0%8F%E5%96%AE%E4%BD%8D#cite_note-19 (accessed.

[6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, June 7, 2015 - June 12, 2015, Boston, MA, United states, 2015, vol. 07-12-June-2015: IEEE Computer Society, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 431-440, doi: 10.1109/CVPR.2015.7298965. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2015.7298965

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI 2015, October 5, 2015 - October 9, 2015, Munich, Germany, 2015, vol. 9351: Springer Verlag, in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp. 234-241, doi: 10.1007/978-3-319-24574-4_28. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-24574-4_28

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in 3rd International Conference on Learning Representations, ICLR 2015, May 7, 2015 - May 9, 2015, San Diego, CA, United states, 2015: International Conference on Learning Representations, ICLR, in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

[9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in 3rd International Conference on Learning Representations, ICLR 2015, May 7, 2015 - May 9, 2015, San Diego, CA, United states, 2015: International Conference on Learning Representations, ICLR, in 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in 29th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, June 26, 2016 - July 1, 2016, Las Vegas, NV, United states, 2016, vol. 2016-December: IEEE Computer Society, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 770-778, doi: 10.1109/CVPR.2016.90. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.90

[11] A. Dosovitskiy et al., "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE," in 9th International Conference on Learning Representations, ICLR 2021, May 3, 2021 - May 7, 2021, Virtual, Online, 2021: International Conference on Learning Representations, ICLR, in ICLR 2021 - 9th International Conference on Learning Representations, p. Amazon; DeepMind; et al.; Facebook AI; Microsoft; OpenAI.

[12] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 4, pp. 834-848, 2018, doi: 10.1109/TPAMI.2017.2699184.

[13] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," arXiv, 2018.

[14] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze and excitation' in fully convolutional networks," arXiv, 2018.

[15] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," arXiv, 2017.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019, June 2, 2019 - June 7, 2019, Minneapolis, MN, United states, 2019, vol. 1: Association for Computational Linguistics (ACL), in NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, pp. 4171-4186.

[17] J. Chen et al., "TransUNet: Transformers make strong encoders for medical image segmentation," arXiv, 2021.

The End

Thank you for your attention

附錄

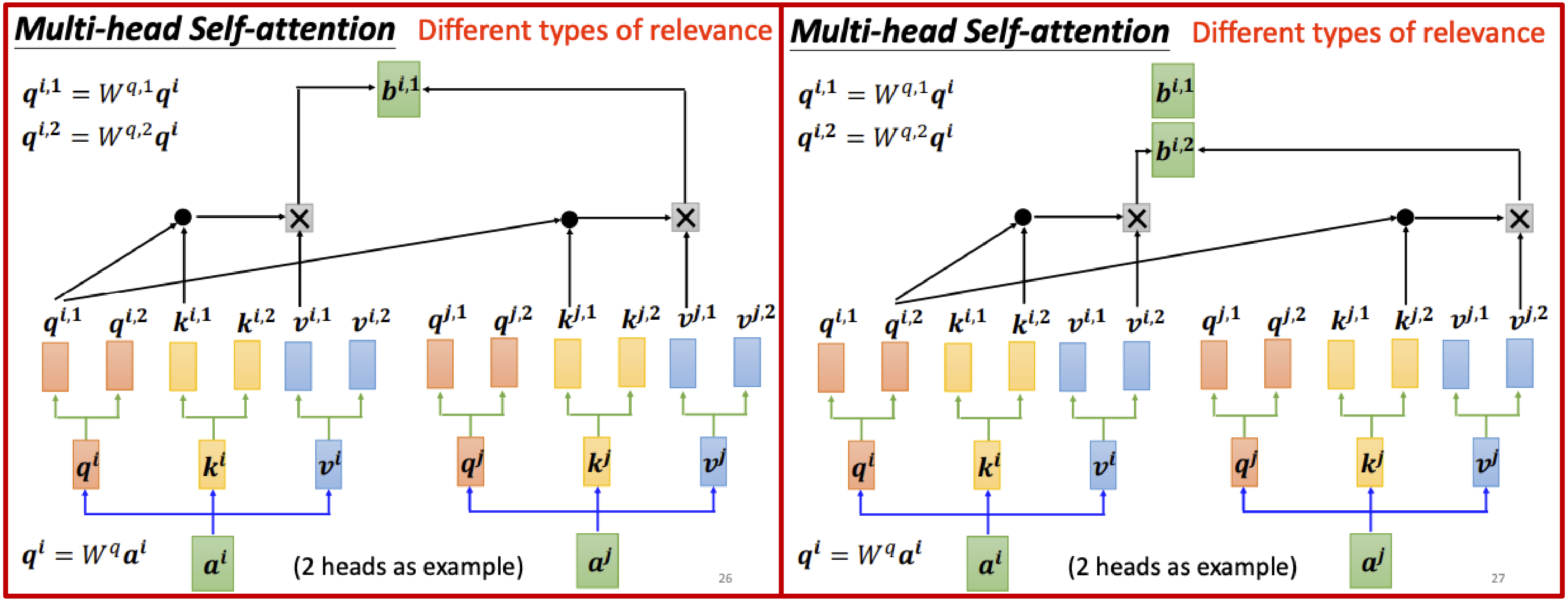
Global frameworks — Interpolate

F.interpolate(x, scale_factor=2, mode= “bilinear”)

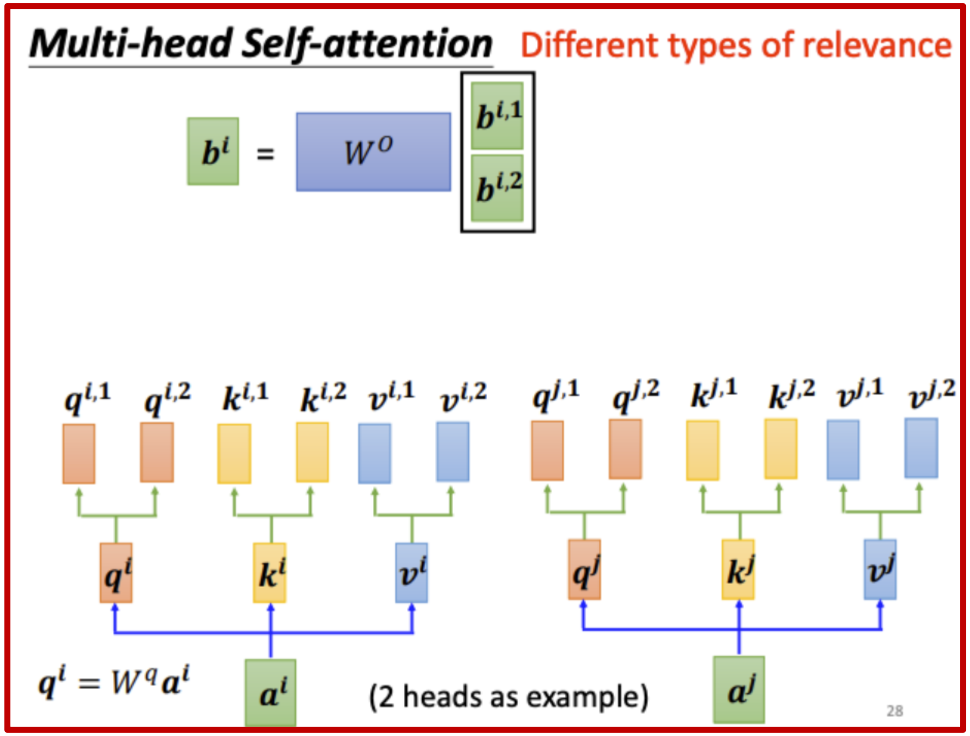
```
tensor([[[[1., 2.],
          [3., 4.]]]])
```

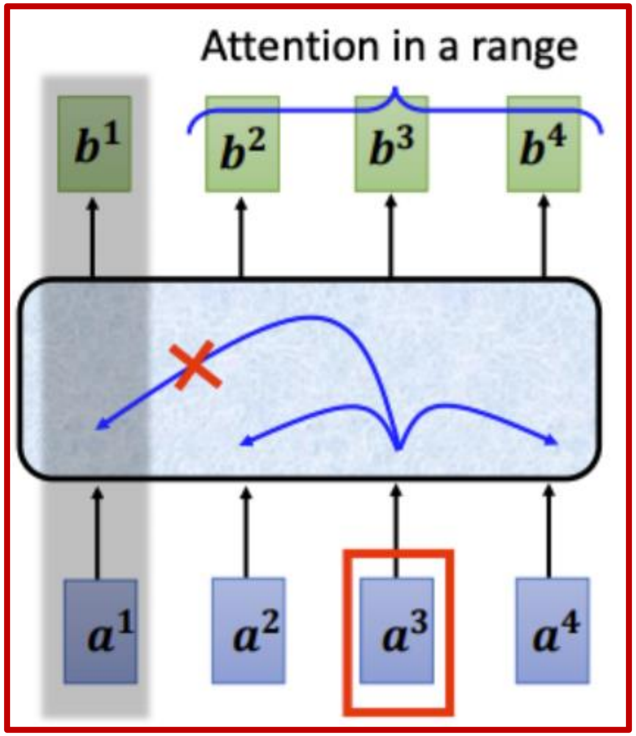
```
tensor([[[[1.0000, 1.3333, 1.6667, 2.0000],
          [1.6667, 2.0000, 2.3333, 2.6667],
          [2.3333, 2.6667, 3.0000, 3.3333],
          [3.0000, 3.3333, 3.6667, 4.0000]]]])
```

Multiheaded Self-attention (MSA)



Multiheaded Self-attention (MSA)





Confusion matrix

IoU thredshold=0.5